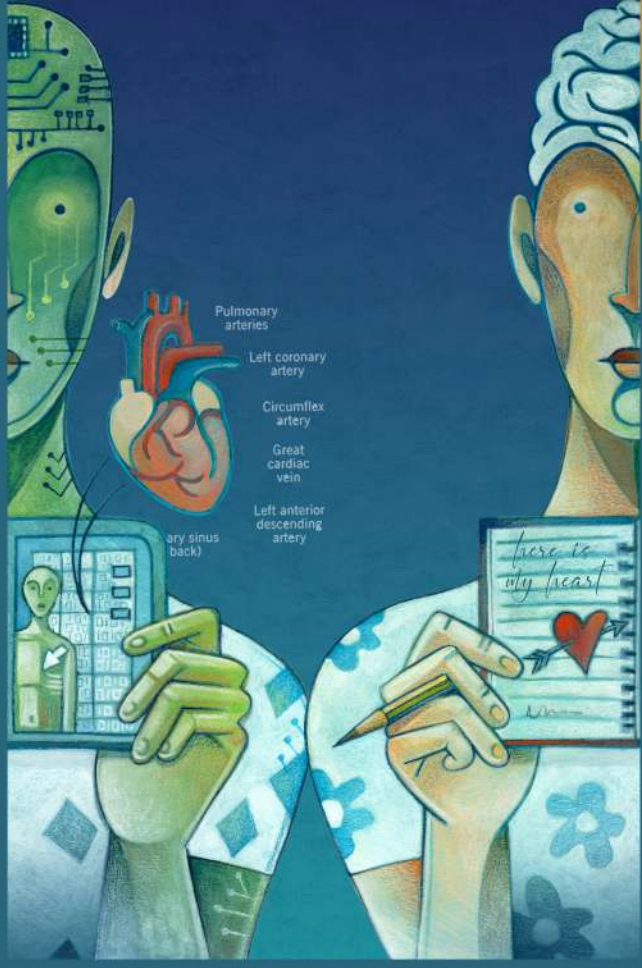
Lorella Giannandrea
Francesca Gratani
Lorenza Maria Capolla
[a cura di]



VALUTAZIONE E INTELLIGENZA ARTIFICIALE

Prospettive didattiche, etiche e metodologiche

sentieri scenari postmoderni

Pensa MULTIMEDIA

sentieri
scenari
postmoderni

Collana diretta da
Michele Corsi
Massimiliano Stramaglia

12

Lorella Giannandrea, Francesca Gratani
Lorenza Maria Capolla
[a cura di]

VALUTAZIONE E INTELLIGENZA ARTIFICIALE

PROSPETTIVE DIDATTICHE, ETICHE E METODOLOGICHE

ATTI DEL CONVEGNO PRIN "AI&F - ARTIFICIAL INTELLIGENCE & FEEDBACK FOR EFFECTIVE LEARNING".
22 E 23 GENNAIO 2026
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Pensa MULTIMEDIA

Il presente lavoro è stato pubblicato con i fondi del Progetto PRIN 2022
AI&F - Artificial intelligence & feedback for effective learning
n. 2022ZMYYTH - CUP D53D23013110006 - PNRR M4.C2.1.1

ISBN volume 979-12-5568-510-4
ISSN collana 2421-1133

2026 © by Pensa Evolution Srl
73100 Lecce • Via Arturo Maria Caprioli, 8 • Tel. 0832.230435
www.pensamultimedia.it

SOMMARIO

Introduzione

Lorella Giannandrea
Università di Macerata
Francesca Gratani
Università di Macerata
Lorenza Maria Capolla
Università di Macerata

Negli ultimi anni, il rapporto tra valutazione e apprendimento è stato al centro del dibattito internazionale, sollecitato da trasformazioni profonde che riguardano tanto le pratiche didattiche quanto gli strumenti disponibili per sostenerle.

Una visione della valutazione come pratica prevalentemente certificativa, che si colloca al termine dell'azione didattica per esaminare i risultati, è stata affiancata dall'idea di una valutazione come processo utile a migliorare l'insegnamento e l'apprendimento. Studi ormai consolidati, infatti, hanno messo in evidenza come la valutazione e il feedback rappresentino leve potenti per sostenere l'apprendimento purché progettati come parte integrante delle attività didattiche e non come un momento successivo o accessorio (Hattie & Clarke, 2018; Carless & Boud, 2018; Winstone & Carless, 2019). Il paradigma del *learning-oriented assessment* raccoglie questa prospettiva, enfatizzando la centralità dei processi di interpretazione, uso e rielaborazione del feedback da parte degli studenti (Carless, 2015; Laici, 2021).

In questo scenario, la diffusione dell'Intelligenza Artificiale (IA), e in particolare dell'IA generativa, ha contribuito a ridefinire il dibattito sulla valutazione (Ajawi et al., 2023; Bearman & Ajawi, 2023). L'IA non è stata, infatti, un semplice strumento che si è aggiunto a quelli già esistenti, ma ha agito come un fattore di trasformazione capace di rimettere in discussione assunti consolidati, pratiche sedimentate e modelli impliciti di valutazione (Bearman et al., 2024).

Il progetto PRIN *AI&F - AI and Feedback for Effective Learning* si inserisce in questo spazio di tensione. Il percorso di ricerca, avviato nel 2023 dall'Università degli Studi di Macerata in collaborazione con le Università di Bari e Padova, si è posto l'obiettivo di indagare il potenziale dell'IA nel supportare i docenti nell'analisi di compiti aperti o poco strutturati e nella produzione di fe-

edback tempestivi e generativi, in contesti di *higher education* caratterizzati da elevata numerosità di studenti (Foschi et al., 2025). Si è dunque lavorato allo sviluppo di un approccio metodologico fondato sull'interazione tra agente umano e agente artificiale per la costruzione del processo valutativo, unitamente alla definizione di un'interfaccia intuitiva volta a sostenere l'attivazione di *feedback loop* tra docenti e studenti. In linea con gli approcci della *design-based implementation research* (Fishman et al., 2013), il progetto ha privilegiato la sperimentazione e la riprogettazione dei dispositivi educativi in contesti reali e complessi, evitando semplificazioni che avrebbero rischiato di ridurre la complessità dei fenomeni osservati. Prendendo l'avvio dall'analisi dei modelli e delle pratiche valutative adottate negli atenei italiani, si è tentato di individuare approcci ricorrenti, problematiche e possibili linee di sviluppo. Successivamente è stato avviato un lavoro di progettazione, sviluppo e sperimentazione di dispositivi per facilitare i processi di feedback, specialmente in situazioni caratterizzate da gruppi numerosi di studenti e da compiti aperti, esplorando il contributo dell'IA in tale direzione.

Il convegno finale del progetto, svoltosi a Bari il 22 e 23 gennaio 2026, è stato un momento di confronto e di rilancio. I contributi presentati restituiscono la trama di un percorso di ricerca che ha coinvolto più contesti universitari, numerosi insegnamenti e ampie coorti di studenti, mettendo alla prova ipotesi teoriche e soluzioni operative in situazioni autentiche. Accanto ai risultati dei gruppi di ricerca coinvolti nel PRIN *AI&F*, il presente volume raccoglie contributi di studiosi italiani e internazionali che, da prospettive differenti, indagano il rapporto tra valutazione e IA.

Un elemento trasversale ai diversi contributi riguarda il ripensamento della valutazione come problema di progettazione (*assessment as design*). In continuità con le prospettive che concepiscono l'insegnamento come una scienza della progettazione (Laurillard, 2012), la valutazione viene qui intesa come un dispositivo progettuale che struttura le condizioni dell'apprendimento, orienta le attività degli studenti e rende visibili i criteri di qualità. In questo quadro, l'introduzione dell'IA non sostituisce il lavoro progettuale, ma ne evidenzia la centralità. La valutazione emerge, infatti, come un campo di negoziazione continua, in cui entrano in gioco dimensioni epistemologiche, didattiche ed etiche. L'adozione di strumenti di IA non elimina le criticità, ma anzi può renderle più visibili, portando alla luce zone di indeterminazione e incongruenze spesso implicite nelle pratiche valutative. È in questo spazio che l'IA assume una funzione specifica: non tanto fornire soluzioni o sostituire il lavoro e il giudizio umano, quanto sostenere e collaborare come *sparring partner* in processi di esplicitazione e di riflessività. L'interazione tra docenti, studenti e sistemi di IA, così come emerge dalle

ricerche presentate, interviene fin dalle prime fasi del progetto come un processo di co-costruzione dei criteri di valutazione, in cui la definizione dei livelli, l'interpretazione delle risposte e l'attribuzione di valore diventano oggetto di confronto e revisione tra i diversi attori umani e tra agenti umani e artificiali.

Un secondo elemento riguarda il ruolo del feedback. Nei contributi raccolti, esso è concepito come un elemento strutturale del processo di apprendimento. Il feedback, infatti, non consiste semplicemente nella trasmissione di informazioni o nella restituzione di un giudizio, ma si struttura come un processo dialogico e generativo, che richiede il coinvolgimento e la riflessione critica di docenti e studenti (Winstone & Carless, 2019; Laici, 2021; Pentucci, 2022). Si inserisce così in un modello di valutazione distribuita, caratterizzato da una pluralità di occasioni integrate nei percorsi didattici (Giannandrea, 2022) e orientate a sostenere processi di regolazione e autoregolazione (Panadero et al., 2019). L'intelligenza artificiale può ampliare le possibilità di generazione e articolazione del feedback, rendendo possibile un suo utilizzo più tempestivo, continuo e, in alcuni casi, più sostenibile e accessibile anche in contesti caratterizzati da numeri elevati di studenti.

Un ulteriore elemento riguarda il ruolo degli studenti. Come sottolineato da diversi contributi, essi non sono semplicemente destinatari di feedback, ma partecipano ai processi di interpretazione, discussione e revisione dei propri elaborati (Carless & Boud, 2018). L'interazione con l'IA può attivare forme di riflessione metacognitiva e contribuire allo sviluppo di una maggiore consapevolezza dei criteri di qualità, sostenendo una partecipazione più attiva ai processi valutativi e l'attivazione di *feedback loop* che coinvolgano in fasi successive di interazione, docenti, studenti e agenti artificiali.

Queste potenzialità non si traducono automaticamente in innovazione didattica. Richiedono, al contrario, un lavoro intenzionale di progettazione, che include la definizione di criteri espliciti, la costruzione di rubriche condivise, la selezione di compiti autentici e la strutturazione di ambienti di apprendimento in cui il feedback possa essere effettivamente utilizzato. L'introduzione dell'IA rende ancora più esplicita la necessità di una competenza progettuale e valutativa avanzata da parte dei docenti.

Nel loro insieme, i contributi raccolti in questo volume offrono una lettura articolata e non riduzionistica del rapporto tra valutazione e intelligenza artificiale. Più che proporre modelli definitivi o soluzioni generalizzabili, essi restituiscono la prospettiva di una ricerca in evoluzione, caratterizzata da sperimentazioni, tensioni e aperture.

Il volume si articola in due parti.

La prima, a cura delle tre Unità di ricerca del PRIN *AI&F*, presenta il percorso

complessivo del progetto, mettendo in evidenza le scelte teoriche e metodologiche e gli esiti delle diverse fasi di implementazione. Particolare attenzione è dedicata al tema dell'interazione tra agenti umani e artificiali nei processi di feedback e alle trasformazioni delle pratiche valutative nei contesti universitari.

La seconda parte del volume raccoglie i contributi presentati nelle sessioni parallele del convegno ed è organizzata in quattro aree tematiche.

La prima area tematica, dedicata all'**IA nella valutazione universitaria**, esplora una pluralità di pratiche e dispositivi attraverso cui l'intelligenza artificiale entra nei processi valutativi in ambito accademico. I contributi analizzano, da prospettive differenti, l'uso della GenAI nella produzione degli elaborati da parte degli studenti, le trasformazioni dei modelli di *assessment*, l'impiego di chatbot per il supporto alla valutazione formativa o per stimolare la riflessione degli studenti sui compiti assegnati. Vengono infine presentate pratiche in cui l'IA viene utilizzata per lo sviluppo di strumenti per la costruzione di rubriche e criteri di valutazione. I contributi mettono in luce le percezioni di docenti e studenti, facendo emergere nuove sensibilità e consapevolezze e stimolando un pensiero critico rispetto al ruolo dell'IA nei processi valutativi universitari.

La seconda area tematica si concentra sull'**IA nella valutazione a scuola**. I contributi indagano le percezioni degli insegnanti, tra aperture e resistenze, e analizzano gli usi che gli studenti fanno degli strumenti di IA generativa. Particolare attenzione è dedicata alle implicazioni didattiche e valutative di tali pratiche, alle questioni di equità e validità, nonché alle potenzialità dell'IA nel supportare processi inclusivi e metacognitivi. I contributi restituiscono un quadro complesso, in cui opportunità e criticità si intrecciano, rendendo evidente la necessità di una formazione specifica e di una più matura competenza valutativa da parte dei docenti.

La terza area tematica affronta il tema dell'**IA nella formazione degli insegnanti**, mettendo in evidenza come la diffusione di questi strumenti richieda un ripensamento delle competenze professionali, in particolare in relazione alla competenza valutativa. I contributi raccolti spaziano da esperienze di formazione iniziale e in servizio a *scoping reviews* che mirano a costruire uno stato dell'arte della ricerca sull'impatto dell'IA nella formazione docente. Vengono proposti modelli e dispositivi per sostenere un uso critico e consapevole dell'intelligenza artificiale nella costruzione della competenza valutativa dei docenti in formazione e in servizio. L'IA emerge non solo come oggetto di apprendimento, ma come mediatore che può supportare processi riflessivi e trasformativi nella costruzione della professionalità docente.

Infine, la quarta area tematica è dedicata al **rapporto tra IA e feedback**, nodo centrale dell'intero progetto PRIN *AI&F*. I contributi approfondiscono il ruolo

dell'intelligenza artificiale nella generazione, articolazione e utilizzo del feedback, con particolare attenzione allo sviluppo della *feedback literacy* degli studenti e dei docenti, all'efficacia percepita del feedback umano e automatizzato e alla progettazione di dispositivi di feedback orientati a sostenere processi di apprendimento e processi di professionalizzazione degli insegnanti. Vengono inoltre presentate esperienze di utilizzo di chatbot e altri strumenti di IA per la costruzione di feedback personalizzati e per il supporto a compiti complessi, evidenziando le potenzialità di questi dispositivi nel promuovere forme più continue, dialogiche e formative di restituzione.

In conclusione, il progetto *AI&F* invita a superare una visione strumentale dell'intelligenza artificiale per collocarla all'interno di una riflessione più ampia sulla valutazione come pratica educativa. In un contesto in cui le tecnologie evolvono rapidamente, la sfida non è tanto quella di inseguire le innovazioni, quanto di sviluppare cornici interpretative e progettuali capaci di orientarne l'uso in modo critico e consapevole (Corbin et al., 2025; Corbin et al., 2026).

Le diverse prospettive presentate suggeriscono che l'incontro tra valutazione e intelligenza artificiale possa rappresentare un'occasione per ripensare il significato stesso del valutare: non solo misurazione o classificazione, ma processo di costruzione condivisa di senso, in cui criteri, evidenze e giudizi si definiscono attraverso processi dialogici, situati e continuamente rinegoziati. L'IA non viene assunta acriticamente come un dispositivo per automatizzare il giudizio, ma come un attore del sistema capace di amplificare le domande, rendere visibili le ambiguità, sollecitare forme più approfondite di argomentazione e di responsabilità epistemica.

La presenza dell'IA nei contesti educativi non segna, dunque, la fine della valutazione come pratica umana ma, al contrario, ne potrebbe evidenziare la natura profondamente ermeneutica e dialogica. Proprio perché l'IA è in grado di produrre risposte plausibili, ma non definitive, essa induce a esplicitare ciò che spesso resta implicito: i criteri di qualità, le soglie di accettabilità, le visioni di apprendimento che orientano le nostre scelte. Proprio per questo, l'IA può diventare uno strumento per potenziare la riflessività, capace di sostenere docenti e studenti in un lavoro più consapevole sul valore e sul significato dei processi valutativi (Winstone et al., 2026).

Guardando agli sviluppi futuri, il percorso avviato dal progetto *AI&F* apre la possibilità di costruire ecosistemi educativi in cui umano e artificiale non si contrappongano, ma agiscano congiuntamente nella produzione di conoscenza e nella regolazione dei processi di apprendimento.

I contributi qui raccolti indicano, infine, nuovi spazi di ricerca e suggeriscono nuove domande: come cambia il ruolo attribuito alla valutazione nei processi

educativi e come si modificano le posture dei vari attori coinvolti? Le tecnologie possono contribuire a rendere i processi valutativi più equi e trasparenti? Nella creazione degli artefatti, tanto nella ricerca quanto nel lavoro degli studenti, come cambiano le relazioni tra conoscenza, autorialità e interpretazione quando parte del processo è co-mediata da sistemi di IA generativa?

Bibliografia

Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2023). From authentic assessment to authenticity in assessment: Broadening perspectives. *Assessment & Evaluation in Higher Education, 49*(4), 499–510. https://doi.org/10.1080/02602938.2023.2154286

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, 54 (5), 1160–1173.

Bearman, M., Ajjawi, R., Boud, D., Dawson, P., & Tai, J. (2024). *Re-imagining assessment in a digital world*. Springer.

Carless, D. (2015). *Excellence in University Assessment: Learning-oriented Assessment in Action.* Routledge.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education*, 50(7), 1087–1097. https://doi.org/10.1080/02602938.2025.2503964

Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2026). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*, 51(4), 736–752. https://doi.org/10.1080/02602938.2025.2553340

Fishman, B.J., Penuel, W.R., Allen, A., & Cheng, B.H. (Eds.). (2013). *Design-based implementation research: Theories, methods, and exemplars*. New York: Teachers College Record.

Foschi, L.C., Doria, B., Laici, C., Gratani, F., Screpanti, L., Fiorentino, M.G., Rossi, P.G., Giannandrea, L., Grion, V., & Montone, A. (2025). AI e feedback. Interazione tra agenti umani e artificiali per valutare prove scritte in ambito universitario. *Education Sciences & Society, 15*(2).

Giannandrea, L. (2022). La valutazione diffusa: Gli embedded tasks e l'assessment as learning. In P.C. Rivoltella & P.G. Rossi (Eds.), *Nuovo agire didattico* (pp. 165–170). Brescia: Morcelliana-Scholè.

Hattie, J., & Clarke, S. (2018). *Visible learning: Feedback*. Routledge.

Laici, C. (2021). *Il feedback come pratica trasformativa nella didattica universitaria*. Milano: FrancoAngeli.

Panadero, E., Broadbent, J., Boud, D., & Lodge, J.M. (2019) Using formative assessment to influence self- and co-regulated learning: the role of evaluative judgement. *European Journal of Psychology of Education,* 34, 535–557. https://doi.org-/10.1007/s10212-018-0407-8

Pentucci, M. (2022). Il feedback. In P.C. Rivoltella & P.G. Rossi (Eds.), *Nuovo agire didattico* (pp. 101–108). Brescia: Morcelliana-Scholè.

Laurillard, D. (2012). *Teaching as a design science: Building pedagogical patterns for learning and technology*. Routledge.

Winstone, N. E., & Carless, D. (2019). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge.

Winstone, N. E., Gravett, K., & Elkington, S. (2026). Black Box Assessment: rethinking integrity and learning for a time of Generative AI. *Assessment & Evaluation in Higher Education*, 1–18. https://doi.org/10.1080/02602938.2026.2661365

I PARTE

Interventi dei Componenti delle Unità di ricerca del PRIN AI&F

I.1
Interazione tra agenti umani e artificiali per il Feedback Loop: evidenze sperimentali emerse nel PRIN "AI&F"

Pier Giuseppe Rossi[*]
Professore onorario, Università degli Studi di Macerata, piergiusepperossi1@gmail.com
Lorella Giannandrea
Professoressa ordinaria, Università degli Studi di Macerata, lorella.giannandrea@unimc.it
Chiara Laici
Professoressa associata, Università degli Studi di Macerata, chiara.laici@unimc.it
Francesca Gratani
Ricercatrice, Università degli Studi di Macerata, f.gratani@unimc.it
Lorenza Maria Capolla
Assegnista di ricerca, Università degli Studi di Macerata, l.capolla@unimc.it

Abstract

Il contributo indaga il problema della sostenibilità del feedback nei contesti universitari caratterizzati da elevata numerosità, mantenendone al contempo qualità e funzione formativa. Nell'ambito del progetto PRIN "AI&F", lo studio analizza come l'intelligenza artificiale possa supportare i docenti nell'analisi di compiti aperti e nella produzione di feedback tempestivi e generativi, utili ad attivare processi di feedback loop. Il lavoro ricostruisce inoltre l'evoluzione del progetto alla luce dei recenti sviluppi dell'IA, evidenziando come il passaggio a sistemi generativi abbia modificato strumenti e modalità operative, incidendo anche sugli obiettivi e sulle metodologie di ricerca. I risultati mostrano che l'efficacia del feedback dipende dalla progettazione dell'interazione tra docente e sistema, con rilevanti implicazioni per la qualità e la scalabilità dei processi valutativi.

Parole chiave: Machine learning; intelligenza artificiale generativa; feedback; valutazione; università.

* Pier Giuseppe Rossi e Lorella Giannandrea hanno supervisionato la conduzione della ricerca nell'ambito del progetto PRIN2022 e la stesura complessiva dell'articolo. Nello specifico, si segnalano le seguenti attribuzioni: Pier Giuseppe Rossi paragrafi 4.1.1 e 4.1.2; Lorella Giannandrea paragrafo 5; Chiara Laici paragrafo 3; Francesca Gratani paragrafi 2, 4.1.3 e 4.2; Lorenza Maria Capolla paragrafi 1, 4.1.4, 4.1.5 e 4.3.

The paper examines the issue of feedback sustainability in higher education contexts characterized by large student cohorts, while preserving its quality and formative function. Within the PRIN "AI&F" project, the study investigates how artificial intelligence can support teachers in the analysis of open-ended tasks and in the generation of timely, generative feedback capable of activating feedback loops. It also reconstructs the evolution of the project in light of recent advances in AI, showing how the shift to generative systems has reshaped tools and operational practices, with implications for research aims and processes. The findings indicate that the effectiveness of feedback depends on the design of the interaction between the teacher and the system, with significant implications for the quality and scalability of assessment practices.

Keywords: Machine learning; generative AI; feedback; assessment; higher education.

1. Introduzione

La letteratura scientifica ha ampiamente evidenziato il ruolo del feedback nei processi di valutazione e apprendimento (Winstone & Carless, 2019; Lipnevich & Panadero, 2021). Tali evidenze si confrontano tuttavia con un problema ricorrente di sostenibilità: nei contesti caratterizzati da un numero elevato di studenti, risulta spesso difficile garantire feedback tempestivi, pertinenti e realmente utili ai processi di apprendimento. In questa direzione, la ricerca educativa sta progressivamente esplorando soluzioni basate sull'integrazione tra analisi umana e intelligenza artificiale, con l'obiettivo di rendere il feedback più sostenibile senza ridurne il valore pedagogico (Deeva et al., 2021; Gao et al., 2024).

In questo quadro si colloca il progetto PRIN AI&F: *Artificial Intelligence and Feedback for Effective Learning*, finalizzato a indagare se e in che modo l'intelligenza artificiale possa supportare i docenti nell'analisi di un numero elevato di elaborati aperti e nella produzione di feedback tempestivi e generativi. Il progetto assume come riferimento una concezione del feedback non più inteso come trasmissione unidirezionale di un giudizio, ma come processo dialogico attraverso cui lo studente può monitorare, comprendere e riorientare il proprio apprendimento (Sadler, 2010; Winstone & Carless, 2019).

Lo sviluppo del progetto si è intrecciato con una trasformazione particolarmente rapida del campo dell'intelligenza artificiale. Tra il 2022 e il 2026, l'evoluzione dei sistemi disponibili ha modificato non solo gli strumenti utilizzabili, ma anche le condizioni stesse della ricerca. Le finalità generali del PRIN sono rimaste stabili; sono invece mutati, anche in modo significativo, le modalità

operative, alcuni obiettivi intermedi e le forme di interazione tra agente umano e agente artificiale. In particolare, il passaggio da strumenti di *machine learning* che richiedevano competenze specialistiche a sistemi generativi dialogici accessibili in linguaggio naturale ha ridefinito il rapporto tra progettazione didattica, valutazione e feedback.

Il progetto, concepito nel 2021, presentato nel 2022 e avviato operativamente nel 2023, nasce infatti in un contesto tecnologico diverso da quello in cui si è sviluppato. Le tecnologie inizialmente previste appartenevano alla precedente generazione di strumenti di *machine learning*; l'emergere dell'AI generativa (GenAI) ha successivamente consentito di perseguire le stesse finalità con modalità più accessibili e, in parte, più affidabili. Questo slittamento non ha avuto soltanto una portata tecnica, ma ha prodotto una riconsiderazione del processo valutativo e del ruolo dell'interazione tra docente e tecnologia.

Il presente contributo ricostruisce tale evoluzione e presenta i principali risultati emersi nel corso delle sperimentazioni. In coerenza con le prospettive della *Design-Based Implementation Research* (DBIR) (Fishman et al., 2013), l'obiettivo non è soltanto documentare esiti empirici, ma mostrare come la trasformazione degli strumenti abbia inciso sulla definizione degli obiettivi, sulle procedure di valutazione e, più in generale, sui modi di condurre la ricerca educativa in contesti segnati da innovazioni rapide e non lineari. In questa prospettiva, il paper mette in evidenza il ruolo centrale dell'interazione tra agenti umani e artificiali nella costruzione del giudizio valutativo e nell'attivazione del *feedback loop*.

2. Background teorico

Il feedback rappresenta uno dei dispositivi pedagogici più rilevanti nei processi di insegnamento-apprendimento nell'istruzione superiore, incidendo sull'autoregolazione degli studenti, sulle performance accademiche e sulle scelte progettuali del docente.

Negli ultimi decenni si è assistito a un rilevante mutamento paradigmatico: da una concezione del feedback come trasmissione unidirezionale di informazioni (*feedback as telling*) a una visione processuale, dialogica e generativa (*feedback as process*) (Winstone & Carless, 2019). In questa prospettiva, il feedback si configura come un processo iterativo e bidirezionale, che richiede coinvolgimento attivo, riflessione critica e traduzione operativa in strategie di miglioramento (Laici, 2021; Hooda et al., 2022; Pentucci, 2023).

All'interno di tale cornice, il concetto di *Feedback Loop* assume un ruolo centrale. Il feedback non esaurisce la propria funzione nella restituzione di un giudi-

zio, ma acquista valore nella misura in cui viene utilizzato dallo studente per orientare azioni successive, attraverso una sequenza ciclica di monitoraggio, revisione e rielaborazione. L'efficacia del *feedback loop* dipende dalla qualità del design didattico, dall'allineamento con gli obiettivi formativi e dallo sviluppo della *feedback literacy*, intesa come capacità di comprendere, valutare e utilizzare attivamente il feedback (Carless & Boud, 2018; Zhan et al., 2025). In questo senso, la tempestività e la comprensibilità del feedback risultano condizioni decisive per attivare processi metacognitivi e di riorientamento dell'apprendimento.

Ne deriva un ripensamento del ruolo del feedback nella progettazione didattica: da atto conclusivo diventa pratica distribuita e socio-materiale, che agisce all'interno di un ecosistema complesso in cui l'apprendimento emerge dall'interazione tra attori umani e non-umani (Gravett, 2022). In contesti caratterizzati da elevata numerosità e crescente eterogeneità, emergono con forza le questioni della qualità, della sostenibilità e della scalabilità del feedback.

Prima dell'avvento della GenAI, le tecnologie educative avevano già ampliato la rapidità e la diffusione del feedback, ma con limiti evidenti nella gestione di compiti complessi e aperti. L'introduzione dei *Large Language Models* ha modificato qualitativamente lo scenario, rendendo possibile una produzione di feedback più personalizzata, contestualizzata e orientata ai processi (Costa & Murphy, 2025; Zhan et al., 2025). Tuttavia, la GenAI non rappresenta una semplice evoluzione tecnologica, ma implica una riconfigurazione dei ruoli e delle pratiche educative, aprendo alla costruzione di modelli ibridi che combinano scalabilità tecnologica ed *expertise* pedagogica (Bearman & Ajjawi, 2023).

In questo quadro, il focus non può essere posto esclusivamente sulle prestazioni della tecnologia, ma deve essere spostato sul processo attraverso cui il feedback viene costruito. L'agente artificiale (AA) non produce risultati in modo autonomo, ma riutilizza – e rielabora – le informazioni, i criteri e le intenzioni fornite dall'agente umano (AU). Il risultato, pertanto, ma emerge progressivamente nel corso dell'interazione tra AU e AA. In assenza di tale interazione, il feedback generato tende a risultare generico, poco coerente con il contesto e scarsamente utile sul piano cognitivo e metacognitivo.

Diventa pertanto centrale interrogarsi su come progettare l'interazione tra AU e AA: come co-costruire rubriche, come formulare richieste, come strutturare il dialogo. La qualità del feedback dipende infatti dall'allineamento tra le logiche del docente e quelle del sistema, nonché dalla capacità di evitare che l'uso dell'AI produca semplificazioni riduttive della complessità educativa (Corbin et al., 2025; Winstone et al., 2025; Costa & Murphy, 2025).

Dal punto di vista tecnico, i sistemi di AI operano attraverso il riconoscimento di pattern e il calcolo di similarità tra rappresentazioni linguistiche. Nei

modelli precedenti, tale similarità era definita come distanza nello spazio vettoriale tra *embedding* testuali. Nei modelli generativi, essa emerge da un processo interpretativo più complesso, che integra contesto, relazioni semantiche e conoscenze pregresse. Questo passaggio segna una differenza rilevante: la similarità non costituisce più un parametro esplicito e stabile, ma è incorporata in un processo inferenziale dipendente dal contesto e dall'interazione. Tale cambiamento migliora le performance, ma introduce possibili *bias*.

Inoltre, le trasformazioni dell'AI negli ultimi anni, a partire dall'introduzione dei modelli *Transformer* (Vaswani et al., 2017) fino allo sviluppo dei sistemi generativi dialogici, hanno reso possibile un'interazione diretta tra docente e sistema in linguaggio naturale. Questo passaggio ha ridotto la necessità di competenze specialistiche di programmazione e di tecnici informatici all'interno del team di ricerca, spostando il focus dalla programmazione alla progettazione dell'interazione, aprendo nuove possibilità, ma anche nuove criticità per la ricerca educativa.

3. Il progetto

Il progetto PRIN *AI&F: Artificial Intelligence and Feedback for Effective Learning*, è stato condotto dall'Università degli Studi di Macerata in collaborazione con le Università di Bari e Padova. L'obiettivo era sviluppare un metodo di lavoro basato sull'interazione tra AU e AA per la costruzione del giudizio valutativo, insieme a un'interfaccia intuitiva in grado di attivare *feedback loop* nell'*higher education*. Il progetto è stato avviato nel novembre 2023 e si è concluso nel febbraio 2026.

In termini generali, il progetto mirava alla costruzione di un sistema integrato capace di rendere sostenibili processi di feedback, anche in presenza di gruppi numerosi e di compiti aperti o poco strutturati (Foschi et al., 2025).

3.1 Obiettivi e contesto

A partire da questo obiettivo generale, la ricerca si è articolata attorno ad alcune linee di sviluppo principali:

– elaborare modalità di analisi e valutazione di risposte aperte attraverso processi interattivi tra AU e AA;
– progettare e implementare *feedback loop* tra docenti e studenti supportati dalle tecnologie;

– migliorare la sostenibilità e la qualità del processo valutativo, rendendo praticabili forme di valutazione continua, formativa e tempestiva;
– rafforzare le competenze valutative dei docenti e la consapevolezza degli studenti rispetto ai propri processi di apprendimento.

Una prima fase del progetto è stata dedicata all'analisi dei modelli e delle pratiche di feedback nelle università italiane, con l'obiettivo di individuare approcci ricorrenti, criticità e possibili linee di sviluppo. I risultati di tale ricognizione, condotta dall'Università di Padova, sono stati formalizzati nel report *Feedback Frameworks & Practices in Higher Education* e hanno costituito la base teorica per le successive fasi progettuali.

A partire dall'anno accademico 2023/2024 è stato quindi avviato il lavoro di progettazione, sviluppo e sperimentazione dei dispositivi e delle procedure oggetto del progetto.

4. WP2: Lo sviluppo

Il lavoro di progettazione, realizzazione e sperimentazione si è articolato in due aree distinte, ma tra loro integrate. La prima riguardava la costruzione di un sistema supportato dall'AI per l'analisi di compiti a risposta aperta e la produzione di giudizi e feedback per gli studenti. La seconda riguardava la realizzazione di un sistema per la gestione del *Feedback Loop*, capace di rendere visibili agli studenti i giudizi e i feedback prodotti, consentire loro di rispondere e permettere al docente di proseguire iterativamente lo scambio.

I due sistemi interagiscono, ma sono autonomi e utilizzano tecnologie differenti. Nel primo, la presenza dell'AI è particolarmente rilevante; nel secondo, non vi è un uso diretto dell'AI, ma l'impatto pedagogico sul processo di apprendimento è altrettanto significativo. I due sistemi sono stati realizzati in momenti distinti del progetto.

Il primo sistema, quello per la valutazione dei testi degli studenti, è stato realizzato con due modalità in fasi successive. Nella prima fase, collocabile tra la fine del 2023 e l'inizio del 2024, si è scelto di utilizzare BERT (*dbmdz/bert-base-italian-cased*). BERT era ritenuto all'epoca più adatto a cogliere le specificità linguistiche e sintattiche dell'italiano in contesti educativi. Il modello è stato integrato in un'applicazione che confrontava segmenti delle risposte degli studenti con *Testi Target* (TT), cioè formulazioni di riferimento predisposte dal docente. Tale fase ha richiesto il supporto di un esperto informatico.

Nella seconda fase, avviata nel corso dell'a.a. 2024/2025, la diffusione della

GenAI ha reso possibile una diversa modalità di interazione con l'AA, meno dipendente dalla mediazione tecnica e più direttamente gestibile dai docenti.

Parallelamente, nel 2024 è stato sviluppato anche il secondo dispositivo per la gestione del *Feedback Loop*.

4.1 Valutare risposte aperte con il supporto dell'AI

4.1.1 La prima fase sperimentale: BERT

La prima fase sperimentale ha riguardato l'utilizzo di BERT per l'analisi di risposte aperte. Mentre il sistema veniva implementato, contemporaneamente veniva sperimentato sul campo. Il corpus iniziale per la sperimentazione era costituito dagli elaborati di una prova svolta da 264 studenti e studentesse del primo anno di Scienze della Formazione Primaria dell'Università degli studi di Macerata.

L'ipotesi iniziale era che l'agente artificiale potesse valutare gli elaborati misurandone la similarità rispetto a una serie di TT. Le prime prove hanno mostrato che, utilizzando come TT risposte degli studenti, era possibile ottenere un livello di affidabilità accettabile solo fornendo un numero molto elevato di esempi. In un primo test, una concordanza soddisfacente tra AU e AA si raggiungeva solo fornendo 160 TT delle 264 risposte degli studenti, ovvero il docente doveva correggere manualmente 160 testi per permettere all'AA di correggerne correttamente altri 100. Il 71% dei casi era coerente con i risultati ottenuti da una valutazione umana. La differenza era al massimo di 1 livello su 10. Il risultato appariva promettente, ma non sostenibile: richiedeva infatti una correzione preliminare troppo ampia da parte del docente.

Questa prima evidenza ha reso necessario un duplice spostamento. Da un lato, si è rivista la valutazione umana di riferimento, passando da una scala con 10 livelli a una scala a 4 livelli. Tale scala è stata utilizzata da tre docenti che hanno effettuato una correzione autonoma dell'intero corpus e poi hanno confrontato i risultati evidenziando quelli non congruenti. Tale operazione ha confermato quanto già noto in letteratura: anche la valutazione umana presenta margini di variabilità, tanto che nel 17% dei casi i valutatori non hanno trovato accordo e le loro valutazioni differivano ancora di un livello. Il risultato della valutazione triangolata è stato utilizzato come riferimento per la successiva sperimentazione. Va sottolineato che la correzione triangolata si discostava in modo sostanziale dalla valutazione effettuata da un singolo docente.

Dall'altro lato, si è iniziato a lavorare non tanto sulla quantità dei TT, quanto sulla loro struttura. Le sperimentazioni hanno mostrato che TT costruiti in modo essenziale, privi di ridondanze, esempi e dettagli accessori, producevano ri-

sultati migliori di TT lunghi o ricavati direttamente dalle risposte degli studenti. Con 16 TT ben costruiti, distribuiti sui quattro livelli, si è ottenuto un risultato già soddisfacente: circa l'87% dei casi si discostava al massimo di un livello, mentre con 48 TT si arrivava intorno al 91%. Questo risultato ha evidenziato l'importanza della qualità dei TT (Rossi et al., 2024).

Un ulteriore passaggio ha riguardato la costruzione dei TT con frammenti tratti dal manuale o dagli appunti del docente, in forme via via più sintetiche. I risultati migliori sono stati ottenuti utilizzando brevi formulazioni riferite ai concetti chiave richiesti dalla domanda. È emerso inoltre che nella valutazione di elaborati molto brevi o molto lunghi l'AA si discostava maggiormente dalla valutazione umana; per questo motivo si è deciso di escludere dal processo automatizzato la correzione di testi la cui lunghezza fosse molto maggiore o minore rispetto alla lunghezza media delle risposte e di affidarla al docente.

Le prove successive hanno permesso di mettere a fuoco tre elementi rilevanti.

In primo luogo, è emerso che i risultati più affidabili si ottenevano utilizzando TT riferiti al quarto livello, cioè al modello di risposta più vicino alle attese del docente. In questo modo, anziché classificare direttamente le risposte in più livelli, risultava più efficace ordinarle in base alla loro distanza da un riferimento "ideale" e successivamente individuare le soglie di passaggio tra i livelli.

In secondo luogo, è risultato più utile analizzare separatamente ciascun nodo concettuale esaminato dalla domanda, anziché confrontare le risposte con un unico testo globale. Questo approccio ha consentito non solo una valutazione più precisa, ma anche di individuare in modo puntuale le carenze degli elaborati, rendendo la successiva costruzione del feedback più mirata.

Infine, è stata confermata l'importanza della costruzione dei TT: testi sintetici, privi di elementi ridondanti e focalizzati sui nuclei concettuali risultano più efficaci, soprattutto quando il numero dei TT è limitato.

Si è così progressivamente passati da una logica di classificazione a una logica di ordinamento. Più che chiedere al sistema di attribuire direttamente un livello, si è rivelato più affidabile farlo lavorare sulla distanza rispetto a formulazioni esemplari del livello più alto e poi distribuire gli elaborati lungo una scala. Tale scelta ha trovato una giustificazione anche teorica: è possibile costruire esempi relativamente stabili di risposte coerenti con i significati attesi, mentre è impossibile anticipare tutte le forme possibili di non coerenza.

Parallelamente, la sperimentazione ha evidenziato anche l'influenza di aspetti apparentemente secondari, come l'uso del maiuscolo, la presenza di elenchi puntati o specifiche strutture sintattiche. Tali elementi potevano alterare la similarità e introdurre *bias*. Si è quindi deciso di normalizzare i testi prima dell'analisi e di affinare progressivamente i TT anche sotto il profilo grammaticale e sintattico.

Da queste prove sono emerse alcune linee guida per la costruzione dei TT: privilegiare l'ordinamento rispetto alla classificazione diretta; costruire TT per singoli nodi concettuali e non per l'intera risposta attesa; utilizzare formulazioni brevi, lineari e prive di elementi accessori; impiegare solo TT coerenti con le richieste della consegna.

Su questa base è stata definita una procedura più stabile. In primo luogo, venivano esclusi dal corpus automatizzato i testi troppo brevi o troppo lunghi. In secondo luogo, per ciascun nodo concettuale venivano predisposti circa dieci TT di livello 4, inizialmente costruiti a partire dal manuale o dagli appunti del docente e, successivamente, raffinati anche sulla base di elaborati degli studenti già individuati come adeguati. Gli elaborati venivano quindi ordinati in base alla similarità media rispetto ai TT di ciascun nodo. A partire da tale ordinamento, il docente analizzava solo alcuni elaborati collocati nelle soglie tra i quartili per definire i punti di passaggio tra i livelli. In questo modo, invece di correggere integralmente l'intero corpus, era sufficiente esaminare un numero ridotto di casi per calibrare le soglie e verificare l'affidabilità del sistema.

Dal punto di vista pedagogico e metodologico, questa procedura ha mostrato che la valutazione con l'AA non può essere considerata una semplice delega, ma va intesa come processo iterativo e dialogico. L'efficacia del sistema non dipende unicamente dal modello utilizzato, ma dalla qualità dei materiali preparati dal docente, dalle scelte di progettazione e dalla progressiva messa a punto dell'interazione AU-AA.

Quanto all'accuratezza finale, il sistema ha restituito una concordanza con la valutazione umana pari al 71% nei casi di corrispondenza esatta (*delta*[1] *0*), al 23% nei casi con uno scarto di un livello (*delta 1*) e al 6% nei casi con uno scarto di due livelli (*delta 2*). Il dato risulta particolarmente significativo se confrontato con la variabilità tra valutazione umana, effettuata da un solo docente, e la valutazione triangolata.

4.1.2 La seconda fase sperimentale: GenAI

Nelle fasi conclusive del progetto (aa.aa. 2024/2025 e 2025/26), la diffusione e la maturazione della GenAI hanno reso opportuna una sua integrazione nel sistema, sia per confrontarne i risultati con quelli ottenuti con BERT, sia per verificare in che misura essa potesse supportare le procedure di valutazione. L'esperienza maturata nella costruzione dei TT si è rivelata comunque fondamentale.

Il cambiamento ha riguardato soprattutto due aspetti. In primo luogo, la Ge-

1 Per *delta* si intende la differenza tra livello assegnato dall'AA e il livello assegnato dall'AU.

nAI analizza e organizza le risposte degli studenti in modo maggiormente contestualizzato. In secondo luogo, e con implicazioni ancor più rilevanti dal punto di vista pedagogico e metodologico, l'introduzione di un'interfaccia dialogica in linguaggio naturale trasforma la modalità di interazione tra docente e agente artificiale. Ne deriva che il confronto con l'AA non riguarda più soltanto la correzione degli elaborati, ma investe già la fase di progettazione della prova, della rubrica e dei criteri valutativi.

Nelle sperimentazioni è stato utilizzato ChatGPT-4 in una prima fase e GPT-5.3 nella seconda. La scelta di ChatGPT è stata determinata dalla disponibilità dei modelli al momento della ricerca e non da una preventiva analisi comparativa con altre soluzioni. Ulteriori studi potranno estendere la sperimentazione ad altri modelli linguistici. Più che sul modello in sé, l'attenzione si è focalizzata sul processo di allineamento progressivo tra AU e AA. L'AI generativa, infatti, non può essere trattata come un correttore autonomo: la qualità delle sue risposte dipende dalla continuità del dialogo, dalla condivisione delle finalità pedagogiche, dalla chiarezza dei criteri e dalla possibilità di affinare nel tempo rubriche, indicatori, livelli ed ancore. In questo senso, il lavoro con la GenAI si è configurato come una forma di co-progettazione valutativa.

Una prima sperimentazione, realizzata nell'autunno 2025, ha mostrato con chiarezza i limiti di un uso non sufficientemente tarato del sistema. In quel caso, dopo aver costruito e discusso una rubrica con l'AA, si è chiesto al sistema di analizzare l'intero corpus di oltre 300 compiti. L'esito ha evidenziato due criticità: da un lato, la correzione non veniva completata poiché, data l'ampiezza del corpus e la presenza di aree di ambiguità nella rubrica, il tempo di elaborazione dell'AA era tale da andare in *timeout*, ovvero forniva risultati relativi solo a parte del corpus; dall'altro, il livello di affidabilità delle valutazioni risultava ancora basso. Questa prima esperienza ha reso evidente che il dialogo con l'AA non può iniziare nel momento della correzione finale, ma deve essere avviato già nella predisposizione della prova e della rubrica.

Da qui è emersa una seconda fase di lavoro, più strutturata, nella quale la rubrica è stata costruita fin dalla progettazione della prova. Per ciascun indicatore sono stati definiti livelli e descrittori. Dopo l'esecuzione del compito, è stato richiesto all'AA di individuare i pattern ricorrenti. In base ad essi AU e AA hanno ridefinito la rubrica e l'hanno applicata a un sottocampione casuale di 20 elaborati, scelta dettata non solo da ragioni di prudenza metodologica, ma anche dalla necessità di verificare se le categorie costruite fossero sufficientemente chiare da risultare operabili per il sistema. L'AA ha analizzato il campione assegnando un livello per ciascun indicatore e segnalando i casi nei quali incontrava maggiori difficoltà interpretative.

L'elemento più rilevante emerso in questa fase è che gli scarti tra AU e AA

non derivavano, nella maggior parte dei casi, da "errori" del sistema in senso stretto, ma da un disallineamento concettuale nella definizione degli indicatori e dei livelli. In alcuni casi il numero dei livelli si è rivelato eccessivo rispetto alla reale possibilità di discriminarli; in altri, l'analisi del campione ha reso necessario introdurre livelli intermedi per intercettare pattern ricorrenti nelle risposte degli studenti. Sono stati inoltre ridefiniti gli stessi indicatori, rendendoli più coerenti con le produzioni effettive.

A partire dalla correzione dei primi 20 compiti, sono state estratte dalle risposte degli studenti alcune formulazioni esemplificative, utilizzate come ancore per completare la rubrica. Questo passaggio ha permesso all'AA di riformulare i livelli in modo più operativo, traducendoli in forme linguistiche e concettuali più facilmente riconoscibili nei testi. In tal senso, la rubrica si è progressivamente trasformata in uno strumento operativo condiviso tra AU e AA.

Diversamente da quanto ipotizzato, il successivo tentativo di associare a ogni risposta un valore di similarità numerica non ha prodotto vantaggi significativi. Nel caso della GenAI, infatti, l'attribuzione dei livelli non si basa su una misura esplicita e stabile di similarità testuale, ma emerge dall'elaborazione contestuale della risposta rispetto ai criteri della rubrica. Per questo motivo, l'informazione numerica sulla similarità si è rivelata poco utile e, al tempo stesso, ha introdotto un aggravio computazionale senza benefici apprezzabili. La strada più efficace è risultata quindi non quella della quantificazione fine della similarità, ma quella del progressivo disambiguamento concettuale della rubrica.

Sulla base del confronto tra AU e AA, la rubrica è stata progressivamente perfezionata fino a renderla più chiara non solo per il sistema, ma anche per i correttori umani. In questo processo è emerso un dato particolarmente interessante: il lavoro richiesto per rendere i criteri non ambigui per l'AA migliorava anche la qualità complessiva del dispositivo valutativo. Una volta completata questa fase di taratura, il sistema ha analizzato l'intero corpus dei compiti, restituendo i risultati in formato tabellare. A questo punto, l'AA indica anche i casi di confine nei quali incontra difficoltà nella valutazione. L'AU analizza i risultati e si focalizza sui casi di confine: se sono situazioni isolate, definisce il livello. Se invece sono situazioni frequenti, li analizza e motiva le proprie scelte che comunica all'AA il quale ripete la valutazione. Nella fase successiva, infine, l'AA completa il giudizio con un feedback, che discende in modo diretto dalla descrizione dei vari livelli della rubrica. La struttura del feedback può essere diversa da compito a compito. Dalle sperimentazioni emergono varie possibili strutture per il feedback:

– una sintesi complessiva della prova;
– una breve restituzione per ciascun indicatore;

– indicazioni di miglioramento;
– eventuali riferimenti puntuali a materiali del corso.

Di volta in volta il docente deve fornire precise indicazioni per la costruzione del feedback.

Nel complesso, la sperimentazione con la GenAI ha mostrato che il punto decisivo non consiste nella semplice sostituzione di BERT con un modello più potente, ma nel mutamento delle modalità operative e del rapporto tra docente e tecnologia. Se nel caso di BERT il lavoro richiedeva la mediazione di un esperto informatico, si concentrava soprattutto sulla costruzione dei TT e sul calcolo della similarità, con la GenAI il docente dialoga direttamente con l'AA e il centro del processo si sposta sulla predisposizione della prova, sulla negoziazione dei criteri, sulla taratura iterativa della rubrica e sulla possibilità di costruire, attraverso il dialogo, un allineamento progressivo tra logiche valutative umane e funzionamento del sistema. Da questa esperienza derivano le linee guida operative presentate nel paragrafo successivo.

4.1.3 Le linee guida per la valutazione con la GenAI

Sulla base delle sperimentazioni condotte, è stato possibile individuare alcune linee guida operative per l'uso della GenAI nella valutazione di elaborati aperti. Il processo si articola in tre fasi.

Fase 1. Progettazione della prova

Il dialogo tra AA e AU inizia nella fase di progettazione e dà il via al processo di allineamento tra i due agenti. In particolare:

1. l'AU condivide e discute con l'AA la progettazione della sessione o delle sessioni di lavoro su cui verterà la prova;
2. l'AU condivide i materiali che utilizzerà con gli studenti, come articoli, pagine di manuale e altri supporti didattici;
3. l'AU esplicita e discute con l'AA le finalità della prova, gli obiettivi e le competenze relativi al percorso;
4. l'AU elabora con l'AA le consegne e le modalità di svolgimento della singola prova;
5. AU e AA costruiscono una prima bozza di rubrica, che contiene gli indicatori principali.

Fase 2. Taratura della rubrica dopo l'esecuzione della prova

La seconda fase è decisiva, perché trasforma una rubrica iniziale in uno strumento operativo per l'AA e per l'AU. I passaggi principali sono i seguenti:

1. l'AA analizza il corpus ed evidenzia pattern ricorrenti per ciascun indicatore;
2. AU e AA rielaborano la rubrica in base ai pattern individuati, ridefinendo indicatori e livelli;
3. l'AA valuta con la rubrica un sottoinsieme limitato di elaborati (circa 15–20) e segnala i casi di confine o quelli per i quali i criteri risultano ambigui;
4. l'AU esamina le attribuzioni, individua le discordanze e le discute con l'AA, argomentando le ragioni del disaccordo;
5. l'AU analizza le situazioni indicate come ambigue e adatta la rubrica (elimina livelli non sufficientemente distinguibili, aggiunge livelli che raccolgono pattern diversi; aggiunge o accorpa indicatori);
6. l'AU seleziona dagli elaborati analizzati risposte intere o segmenti di risposta da utilizzare come ancore della rubrica;
7. l'AA ripete la valutazione del campione con la nuova rubrica arricchita con le ancore;
8. il processo viene reiterato finché non si raggiunge un grado di accordo soddisfacente sui criteri e l'AA non intercetta aree rilevanti di ambiguità.

Fase 3. Valutazione dell'intero corpus

Solo dopo la taratura della rubrica si passa alla valutazione dell'intero insieme degli elaborati. In questa fase:

1. l'AA analizza il corpus sulla base della rubrica condivisa ed evidenzia casi critici; nel prompt relativo a questo passo è opportuno chiedere non solo l'attribuzione dei livelli, ma anche la segnalazione dei casi in cui è difficile attribuire un livello;
2. se emergono frequenti casi di incertezza, è probabile che la rubrica contenga ancora ambiguità. In questi casi è necessario tornare alla fase di taratura;
3. l'AU controlla a campione i risultati (ad esempio, verificando una ragionevole distribuzione dei livelli, la completezza della correzione e la coerenza dell'output);
4. quando la valutazione risulta sufficientemente affidabile, si chiede all'AA di elaborare feedback personalizzati per ogni studente.

Il processo in sintesi

Dopo l'esecuzione del compito, il percorso può essere strutturato nei seguenti passaggi:

– individuazione dei pattern da parte dell'AA;
– taratura della rubrica in base all'analisi di un numero limitato di elaborati;
– analisi dell'intero corpus;
– elaborazione e condivisione dei feedback.

4.1.4 Riflessioni emerse dalla sperimentazione

Le fasi descritte consentono di mettere a fuoco alcuni elementi.

In primo luogo, i criteri di valutazione devono risultare coerenti con le finalità del docente e operativi per le logiche dell'agente artificiale. Le due esigenze non coincidono automaticamente, perché rinviano a modalità differenti di costruzione del significato e richiedono quindi un lavoro di traduzione reciproca. In questa prospettiva, le definizioni descrittive dei livelli non sono sufficienti: l'AI generativa lavora in modo più affidabile quando può confrontarsi con ancore esemplificative, cioè con formulazioni concrete che rendano operativi gli indicatori e i livelli della rubrica.

Un secondo elemento riguarda il funzionamento stesso della GenAI. Il sistema tende infatti a produrre, a "generare", comunque una risposta anche quando i criteri sono ancora ambigui. Per questo motivo il rischio di *bias* aumenta sensibilmente quando il processo viene delegato integralmente o parzialmente all'AA. Prima di procedere all'assegnazione dei livelli è quindi opportuno prevedere almeno due passaggi preliminari: chiedere all'AA di individuare pattern ricorrenti nelle risposte e di segnalare gli elaborati nei quali l'attribuzione del livello risulterebbe ambigua, ovvero quei casi in cui l'assegnazione dei livelli richiederebbe tempi più lunghi. L'assegnazione dei livelli va realizzata solo dopo aver disambiguato i criteri. Se questo lavoro preliminare non viene svolto, l'AI generativa tende comunque a "chiudere" il compito producendo risposte solo apparentemente plausibili, ma spesso poco affidabili, oltre che più onerose in termini di tempo e risorse.

Ne deriva che la valutazione con GenAI non può essere pensata come un atto unico, ma deve essere costruita progressivamente, attraverso richieste parziali, revisioni successive e momenti di confronto tra AU e AA. Solo un processo di questo tipo consente di ridurre i margini di "invenzione" del sistema, che quando non è adeguatamente guidato tende a generare interpretazioni improprie. Il processo permette anche di affinare la rubrica e i casi in cui l'AA segnala ambiguità evidenziano la presenza di elementi di soggettività che possono rendere difficile

la condivisione dei criteri anche tra valutatori umani. In questa prospettiva, il principale vantaggio dell'AI non consiste tanto in un semplice risparmio di tempo, quanto nella possibilità di rendere l'intero processo valutativo più coerente, trasparente e controllabile.

4.1.5 Da BERT alla GenAI

Il confronto con la fase precedente evidenzia alcune differenze rilevanti nel passaggio da BERT all'AI generativa. In primo luogo, cambia la natura dell'interazione: con la GenAI il docente dialoga direttamente con l'agente artificiale in linguaggio naturale, senza la necessità di una mediazione tecnica. Questo spostamento non è solo operativo, ma incide sulla possibilità per il docente di governare il processo.

Cambia il momento in cui l'interazione tra AA e AU si attiva. Se nella fase precedente il confronto con l'AA si collocava prevalentemente nella valutazione degli elaborati, con la GenAI esso si estende alla fase di progettazione del percorso e della prova, contribuendo a costruire fin dall'inizio un allineamento tra intenzioni didattiche e criteri valutativi.

Infine, cambia il modo in cui la similarità viene rappresentata e utilizzata. Nei modelli come BERT, la similarità è una proprietà metrica relativamente stabile, definita dalla distanza tra vettori nello spazio semantico. Nei modelli generativi, invece, essa è il risultato di un processo inferenziale basato sul contesto più sensibile al prompt, al contesto e alla storia del dialogo. La similarità non è quindi un dato fisso, ma una stima che emerge all'interno di una dinamica situata.

Ne deriva che la GenAI offre una maggiore capacità interpretativa, ma al tempo stesso una minore stabilità essendo fortemente legata al contesto. Proprio per questo motivo richiede un lavoro più intenso di allineamento tra AU e AA e una progettazione più accurata delle rubriche e delle interazioni.

4.2 Attivare il feedback loop tra docente e studenti

I risultati della valutazione effettuata con il supporto dell'AI generativa si sono dimostrati complessivamente affidabili; tuttavia, permanevano inevitabili zone di confine in cui il giudizio dell'AA e quello dell'AU potevano in parte divergere. Va sottolineato che tale criticità non è specifica dell'AI: analoghe discrepanze possono emergere tra diversi valutatori umani, o anche nelle valutazioni dello stesso docente effettuate in tempi differenti. Una piena coerenza nella correzione di elaborati aperti è, di fatto, difficilmente raggiungibile. Tuttavia, l'incoerenza risulta generalmente più accettata quando il valutatore è umano.

Il problema pertanto è strutturale e non è risolvibile chiedendo al docente di ricontrollare tutte le valutazioni, operazione peraltro molto dispendiosa in termini di tempo. Per questo motivo, si è ritenuto necessario sviluppare un dispositivo che rendesse il processo valutativo più trasparente e condiviso con gli studenti. La soluzione individuata è stata quella di rendere esplicita, indicatore per indicatore, la logica della valutazione, mostrando rubrica, indicatori e livelli, e di attivare un confronto diretto con gli studenti.

Questa scelta risponde anche a una precisa intenzionalità pedagogica. La possibilità di accedere alla valutazione, di ricevere un feedback puntuale, di commentarlo e di discuterlo con il docente consente allo studente di esplicitare le logiche sottese alla propria risposta e di discuterle con l'insegnante. Inoltre, offre al docente l'opportunità di comprendere in modo più approfondito i processi di pensiero dello studente, fornire ulteriori indicazioni e, se necessario, rivedere il giudizio alla luce di nuovi elementi.

Il feedback, in questa prospettiva, non si configura più come un atto unidirezionale, ma come un processo iterativo, fondato su cicli successivi di interazione tra docente e studente. Per rendere operativa tale impostazione è stato sviluppato un applicativo informatico finalizzato alla gestione del *Feedback Loop*.

L'applicativo *"AIFeel"* consente a ciascun docente di operare all'interno di un'area riservata, nella quale caricare le prove e le relative valutazioni. Elaborati e valutazioni vengono inizialmente inseriti in un foglio elettronico, successivamente condiviso con l'applicativo. Per ogni attività, il docente può dunque servirsi di *AIFeel* per rendere visibili agli studenti gli elaborati, le valutazioni, gli indicatori utilizzati e i feedback. Una volta pubblicati gli esiti, lo studente accede alla propria area riservata tramite link personale, visualizza le proprie risposte, i giudizi e i commenti ricevuti, e può inserire un riscontro in un'apposita chat. Il docente, a sua volta, può rispondere nello stesso spazio, chiarendo, precisando o rilanciando la discussione. In questo modo si attiva un ciclo di interazione potenzialmente infinito, in cui il *Feedback Loop* si realizza concretamente. Il docente può anche decidere di inserire la sua risposta nello stesso foglio elettronico complessivo in cui sono presenti le valutazioni dell'AA, senza accedere alle singole pagine dello studente. La visualizzazione globale delle risposte e dei feedback permette di mantenere sotto controllo la coerenza delle valutazioni e di confrontare agevolmente i diversi interventi, oltre a ridurre i tempi di lavoro.

Dal punto di vista organizzativo, il sistema prevede diversi livelli di gestione (amministrazione di sistema, di ateneo o di dipartimento) e a ogni docente è consentita la visualizzazione solo delle proprie attività. L'applicativo è attual-

mente disponibile per le università interessate ed è stato sperimentato in diversi insegnamenti delle sedi coinvolte nel progetto[2].

Il passaggio dalla prima fase alla seconda fase di sperimentazione, ha modificato la struttura del processo di valutazione. Si presenta ora una sintesi grafica degli step effettuati, sia con il modello BERT che con la GenAI. Il processo sviluppato utilizzando BERT può essere ricondotto a tre step separati (cfr. Fig. 1). Nel primo step il docente predispone la prova, definisce la rubrica valutativa e prepara i TT. Nel secondo step, attraverso un'interazione ricorsiva tra AU e AA, si affinano i TT e si ottiene una classificazione sufficientemente affidabile delle risposte. Nel terzo step, *AIFeel* conclude il percorso valutativo consentendo un processo dialogico tra docente e studenti.

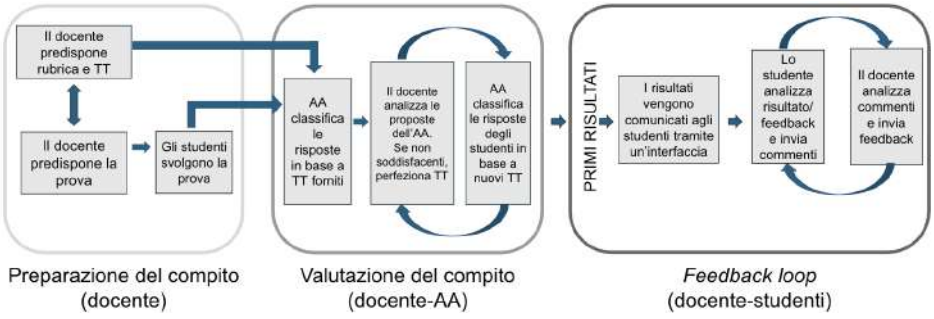


Fig. 1 – Il processo in sintesi (BERT)

L'introduzione di ChatGPT, in sostituzione di BERT, ha comportato una trasformazione significativa del processo. Le diverse *affordances* dell'AI generativa, già discusse nel paragrafo 4.1.2, hanno infatti modificato non solo la valutazione dei testi degli studenti, ma anche la struttura dell'interazione tra docente e AA, che si è spostata da una logica prevalentemente tecnica e parametrica a una logica dialogica e interpretativa, incidendo sia sulla progettazione sia sulla valutazione.

Queste trasformazioni hanno reso necessario un ripensamento complessivo del processo, che assume una configurazione più complessa, come rappresentato nella Fig. 2.

2 Nello specifico, la sperimentazione è stata svolta negli Atenei di Padova, Bari e Macerata, in corsi di laurea di Scienze dell'Educazione, Scienze della Formazione Primaria, Scienze e tecniche psicologiche, Matematica. Complessivamente sono stati coinvolti oltre 1400 studenti.

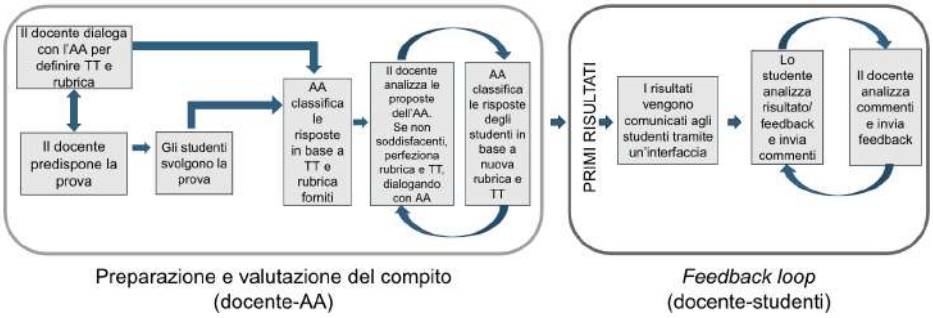


Fig. 2 – Il processo in sintesi (GenAI)

4.3 La fase di testing

La fase di testing è stata realizzata nelle tre sedi coinvolte nel progetto e ha interessato diversi insegnamenti. In una prima ipotesi si prevedeva di testare due sistemi distinti: quello finalizzato all'assegnazione alla valutazione e alla produzione del feedback e quello dedicato alla gestione del dialogo con gli studenti. Tuttavia, l'introduzione dell'AI generativa ha modificato questo assetto, rendendo obsoleto il primo sistema realizzato e orientando il testing verso la validazione delle procedure e delle linee guida. Nel complesso, il testing ha confermato la validità delle linee guida proposte, permettendo al tempo stesso di introdurre alcune variazioni migliorative.

L'attenzione principale si è quindi concentrata sull'applicativo per la gestione del *Feedback Loop*, e i risultati emersi risultano coerenti con quanto già osservato nella fase di messa a punto. In particolare, il sistema consente di gestire in modo rapido ed efficace il dialogo tra docente e studenti.

I riscontri degli studenti sono stati nel complesso positivi. La valutazione pedagogica del processo non si limita alla percezione di trasparenza o alla possibilità di "controllare" il giudizio attraverso il riferimento esplicito alla rubrica. I docenti hanno rilevato anche effetti più profondi, legati al miglioramento dei processi di apprendimento: una maggiore consapevolezza delle proprie conoscenze, una più esplicita rielaborazione delle risposte e una partecipazione più attiva al processo valutativo. Tali evidenze risultano coerenti con quanto riportato nella letteratura internazionale sul feedback, richiamata nella parte iniziale del contributo.

Quando i giudizi e feedback sono restituiti con tempestività e regolarità, il ciclo di interazione può incidere in modo significativo sui processi di insegnamento-apprendimento. Per lo studente, esso consente di rielaborare a posteriori

la prova, riflettere sulle proprie scelte, acquisire maggiore consapevolezza dei criteri valutativi, individuare errori e possibili direzioni di miglioramento e, soprattutto, partecipare attivamente al processo valutativo. Per il docente, il *Feedback Loop* rappresenta uno strumento per esplicitare la logica delle proprie valutazioni, confrontarsi con il punto di vista degli studenti e raccogliere elementi utili per riprogettare attività didattiche e prove successive, fino a co-definire l'esito finale insieme allo studente.

Tali aspetti risultano particolarmente rilevanti in contesti caratterizzati da elevata numerosità, come nel caso delle sperimentazioni condotte, spesso con circa 300 studenti, nei quali sarebbe altrimenti difficile garantire un'interazione individuale significativa.

Accanto ai benefici, sono tuttavia emerse alcune criticità. In particolare, gli studenti tendono a interagire maggiormente in presenza di esiti negativi, mentre sono meno interessati al confronto a fronte di valutazioni positive.

Questi elementi sollecitano alcune implicazioni rilevanti. In primo luogo, l'attivazione di pratiche riconducibili all'*assessment as learning* richiede la predisposizione di uno spazio-tempo strutturato e intenzionalmente dedicato, nonché una postura docente riflessiva e un adeguato accompagnamento degli studenti nello sviluppo di competenze autovalutative. L'efficacia del *Feedback Loop* risulta quindi strettamente connessa alla sua integrazione nel design didattico, non come momento accessorio o finale, ma come dispositivo strutturante dell'intero percorso formativo; la regolarità e la continuità della pratica ne rappresentano condizioni essenziali.

Un ulteriore aspetto riguarda il superamento della configurazione prevalentemente diadica del feedback (docente–studente), in favore di una sua riconfigurazione in chiave collettiva. Le criticità emerse a livello individuale risultano infatti spesso trasversali all'intero gruppo classe, e la creazione di spazi di confronto condivisi può favorire processi di co-costruzione del significato, rafforzando la funzione regolativa e trasformativa del feedback.

5. Conclusioni: riflessioni e prospettive future

La difficoltà principale del progetto PRIN, oggetto del presente contributo, è consistita nel mantenere significatività e coerenza rispetto alle finalità della ricerca in un contesto caratterizzato da profonde e rapide trasformazioni, in particolare sul piano tecnologico. Tale condizione ha richiesto non solo una revisione delle modalità operative, ma soprattutto una ridefinizione progressiva degli obiettivi stessi.

In questo quadro, il contributo del progetto non può essere ricondotto al

solo sviluppo di uno strumento tecnologico, ma alla messa a punto di un diverso modo di concepire il processo valutativo. Il problema iniziale, come garantire un feedback al tempo stesso sostenibile e di qualità, non ha trovato risposta in una soluzione tecnica, ma nella ristrutturazione del processo, che integra in modo sistematico l'interazione tra agente umano e agente artificiale.

Il risultato principale della ricerca risiede dunque nella definizione di un processo dialogico AU-AA, fondato su dinamiche iterative di taratura e allineamento progressivo. In tale processo, il giudizio non è prodotto in modo automatico, ma emerge da una costruzione condivisa, nella quale i ruoli di proposta e di controllo vengono alternativamente assunti dai due agenti. La qualità dell'esito dipende da un lavoro asintotico di affinamento, che rende non delegabile il processo valutativo e richiede una partecipazione attiva e riflessiva da parte del docente.

Questo spostamento implica anche un rovesciamento della prospettiva sull'AI. Il nodo non riguarda tanto le prestazioni del sistema, quanto le condizioni che ne rendono significativo l'utilizzo. L'AI non si configura come un soggetto autonomo, ma come un dispositivo che restituisce, rielabora e amplifica le strutture concettuali e operative introdotte dall'agente umano. In questo senso, la qualità dei risultati dipende dall'interazione e dal contesto in cui è inserita. Il problema non è quindi la tecnologia in sé, ma il dispositivo pedagogico e valutativo entro cui essa viene utilizzata.

Le implicazioni sul piano didattico risultano rilevanti. La valutazione tende a configurarsi come un processo esplicito, discutibile e negoziabile, nel quale criteri e giudizi vengono resi trasparenti e possono essere oggetto di confronto. Il feedback si trasforma da prodotto informativo a processo dialogico e iterativo, mentre la rubrica non si presenta come uno strumento dato a priori, ma come un artefatto che si costruisce e si affina nel corso dell'interazione. In tale prospettiva, il processo valutativo assume una configurazione triangolata che coinvolge docente, studenti e AI, favorendo l'attivazione dell'*agency* dello studente e una maggiore responsabilizzazione di tutti gli attori coinvolti. La prospettiva futura è quella di coinvolgere gli studenti nel processo dialogico fin dalla fase di progettazione iniziale e non solo nella fase finale.

Accanto a tali elementi, la sperimentazione ha evidenziato anche alcuni limiti. In primo luogo, il processo risulta fortemente dipendente dal contesto e dalle competenze del docente, rendendo difficile una replicabilità automatica. Inoltre, il tempo richiesto, pur ridotto in alcune fasi, rimane significativo, in quanto il processo necessita di iterazioni ripetute e momenti di calibrazione. Va inoltre considerata la minore stabilità dei sistemi di AI generativa, la cui risposta può

variare in funzione del contesto e della storia dell'interazione, nonché la presenza di questioni aperte legate alla gestione dei dati e agli aspetti etici.

Un'ultima riflessione riguarda il piano metodologico della ricerca. L'evoluzione intervenuta durante il progetto, segnata dall'introduzione della GenAI, ha richiesto una ridefinizione continua di obiettivi, strumenti e procedure. Ciò sollecita una riflessione più ampia sui modelli di ricerca in contesti caratterizzati da innovazioni rapide e non lineari. In tali condizioni, approcci lineari e predittivi mostrano evidenti limiti, mentre risultano più adeguati modelli iterativi, adattivi e riflessivi. In questa prospettiva, il lavoro svolto si colloca in prossimità dei paradigmi della *Design-Based Implementation Research*, nei quali la produzione di conoscenza è intrecciata ai processi di progettazione, sperimentazione e ridefinizione situata dei dispositivi educativi.

In tale quadro, la validità della ricerca non può essere ricondotta esclusivamente alla replicabilità del disegno iniziale, ma va ripensata in relazione alla trasparenza del processo, alla tracciabilità delle scelte e alla coerenza interna tra finalità, strumenti e interpretazioni. Più che ridurre la variabilità del contesto, si tratta di renderla esplicita e di assumerla come componente strutturale della ricerca.

Riferimenti bibliografici

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, 54 (5), 1160-1173.

Carless, D., & Boud, D. (2018). The development of student feedback literacy. *Assessment & Evaluation in Higher Education*, 43 (8), 1315-1325.

Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education*, 50 (7), 1087-1097.

Costa, C., & Murphy, M. (2025). Generative artificial intelligence in education: (what) are we thinking? *Learning, Media and Technology*, 1-12.

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, 162, 104094.

Foschi, L.C., Doria, B., Laici, C., Gratani, F., Screpanti, L., Fiorentino, M.G., Rossi, P.G., Giannandrea, L., Grion, V., & Montone, A. (2025). AI e feedback. Interazione tra agenti umani e artificiali per valutare prove scritte in ambito universitario. *Education Sciences & Society - Open Access*, 15 (2).

Fishman, B.J., Penuel, W.R., Allen, A., & Cheng, B.H. (Eds.) (2013). *Design-based implementation research: Theories, methods, and exemplars*. New York: Teachers College Record.

Gao, R., Merzdorf, H.E., Anwar, S., Hipwell, M.C., & Srinivasa, A. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6, 100206.

Gravett, K. (2022). Feedback literacies as sociomaterial practice. *Critical Studies in Education*, 63 (2), 261–274.

Hooda, M., Rana, C., Dahiya, O., Rizwan, A., & Hossain, M.S. (2022). Artificial intelligence for assessment and feedback to enhance student success in higher education. *Mathematical Problems in Engineering*.

Laici, C. (2021). *Il feedback come pratica trasformativa nella didattica universitaria*. Milano: Franco Angeli.

Lipnevich, A.A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education*, 6, 720195.

Pentucci, M. (2023). Il feedback. In P.C. Rivoltella, P.G. Rossi (Eds.), *Nuovo agire didattico* (pp. 101–107). Brescia: Scholè.

Rossi, P.G., Giannandrea, L., Gratani, F., Scaradozzi, D., & Screpanti, L. (2024). Teacher-AI interaction in the selection of target texts. In *AIxEDU 2024 Artificial Intelligent Systems in Education 2024* (Vol. 3879, pp. 1-15). CEUR Workshop Proceedings.

Sadler, R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35 (5), 535-550.

Winstone, N., & Carless, D. (2019). *Designing effective feedback processes in higher education: A learning-focused approach*. London: Routledge.

Winstone, N.E., Gravett, K., Noble, C., et al. (2025). Manifesto for feedback in the age of generative artificial intelligence. *Copenhagen Feedback Symposium*, Copenhagen, Denmark.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N.,& Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Zhan, Y., Boud, D., Dawson, P., & Yan, Z. (2025). Generative artificial intelligence as an enabler of student feedback engagement: A framework. *Higher Education Research & Development*, 44 (5), 1289–1304.

I.2
Pratiche di feedback e feedback automatizzato nell'istruzione superiore: analisi delle pratiche dei docenti e delle percezioni degli studenti nel progetto PRIN AI&F*

Valentina Grion
Facoltà di Scienze umane, della formazione e dello sport, Università Pegaso
Laura Carlotta Foschi
Dipartimento di Studi Umanistici, Università degli Studi di Trieste
Beatrice Doria
Facoltà di Scienze umane, della formazione e dello sport, Università Pegaso
Giorgia Slaviero
Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA)
Università degli Studi di Padova
Juliana Elisa Raffaghelli
Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA),
Università degli Studi di Padova
Graziano Cecchinato
Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA),
Università degli Studi di Padova
Corrado Petrucco
Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA),
Università degli Studi di Padova

Abstract

Il feedback è un elemento centrale nei processi di apprendimento universitario, ma la sua integrazione sistematica nelle pratiche didattiche, soprattutto attraverso strumenti digitali e sistemi automatizzati, risulta ancora limitata (Doria et al.,

* Le seguenti attribuzioni riguardano la scrittura dell'articolo e la conduzione della ricerca nell'ambito del progetto PRIN 2022: Grion ha supervisionato la stesura complessiva dell'articolo ed elaborato la discussione e conclusione. Grion e Doria hanno elaborato il disegno di ricerca. Doria e Slaviero hanno curato le analisi del primo nucleo di ricerca. Foschi ha curato le analisi e i risultati del secondo nucleo di ricerca e ne ha proposto alcuni elementi da mettere in discussione. Raffaghelli ha realizzato il questionario relativo al secondo nucleo di ricerca. Cecchinato e Petrucco hanno collaborato alla realizzazione della ricerca.

2025a, 2025b). In questo quadro si inserisce il progetto PRIN "Artificial Intelligence & Feedback for Effective Learning (AI&F)", volto ad analizzare e sviluppare pratiche di feedback supportate da tecnologie digitali nell'istruzione superiore. Il contributo presenta i risultati del WP1, articolato in due nuclei: l'analisi di 690 syllabi di docenti appartenenti a tre università italiane (UniPD, UniMC, UniBA), approfondita tramite focus group, e un'indagine sulle percezioni ed esperienze degli studenti mediante questionario. I risultati mostrano una prevalenza di pratiche valutative centrate sul prodotto e un uso limitato di strategie di feedback strutturate e automatizzate. Gli studenti considerano il feedback docente la fonte più utile, pur riportandone un'esperienza discontinua, mentre il feedback automatizzato è percepito come potenzialmente utile ma ancora poco diffuso. Lo studio evidenzia una discrepanza tra il valore attribuito al feedback e la sua effettiva implementazione, sottolineando la necessità di interventi di faculty development e di modelli pedagogici orientati a un'integrazione sostenibile tra feedback umano e tecnologico.

Parole chiave: Feedback; Feedback automatizzato; Faculty development; Intelligenza artificiale; Pratiche valutative.

Feedback is a central element in university learning processes, but its systematic integration into teaching practices, especially through digital tools and automated systems, is still limited (Doria et al., 2025a, 2025b). This is the context for the PRIN project 'Artificial Intelligence & Feedback for Effective Learning (AI&F)', which aims to analyse and develop feedback practices supported by digital technologies in higher education. This paper presents the results of WP1, which is divided into two main areas: the analysis of 690 syllabuses from lecturers at three Italian universities (UniPD, UniMC, UniBA), supplemented by focus groups, and a survey of students' perceptions and experiences using a questionnaire. The results show a prevalence of product-centred assessment practices and limited use of structured and automated feedback strategies. Students consider teacher feedback to be the most useful source, although they report an inconsistent experience, while automated feedback is perceived as potentially useful but still not widely used. The study highlights a discrepancy between the value attributed to feedback and its actual implementation, emphasising the need for faculty development interventions and pedagogical models aimed at sustainable integration between human and technological feedback.

Keywords: Feedback; Automated feedback; Faculty Development; Artificial intelligence; Assessment practices.

1. Framework teorico

Il feedback rappresenta il meccanismo attraverso il quale le e gli studenti ricevono informazioni sulla qualità dei loro prodotti, in riferimento a uno standard di qualità, al fine di ridurre il divario tra le prestazioni attuali e quelle attese, favorendo così cambiamento e miglioramento (Black & Wiliam, 2009; Sadler, 1989). Nel corso del tempo, la ricerca ha progressivamente abbandonato un modello più trasmissivo di feedback per abbracciare una prospettiva socio-costruttivista, che lo concettualizza come processo dialogico attraverso il quale le e gli studenti attivamente costruiscono, monitorano e valutano il proprio apprendimento (Nicol, 2010; Nicol & MacFarlane-Dick, 2006; Price et al., 2010). Questo dialogo può avvenire sia tra docenti e corpo studentesco, sia tra pari, incoraggiando dinamiche di supporto reciproco nell'individuazione di debolezze nell'apprendimento e stimolando processi di analisi e revisione delle conoscenze e di rielaborazione dell'apprendimento (Nicol & MacFarlane-Dick, 2006; Orsmond et al., 2005). Il coinvolgimento di studenti nei processi di feedback sembra promuovere una comprensione e un'applicazione più efficaci delle informazioni ricevute (Grion & Serbati, 2019), nonché l'attivazione di processi di autovalutazione e l'acquisizione di competenze valutative (Doria & Grion, 2020; Nicol, 2010).

Di conseguenza, l'attenzione che tradizionalmente si è concentrata sul contenuto del feedback si sposta verso le condizioni della sua utilità (Evans, 2013; Serbati et al., 2019). Ciò di cui occuparsi risulterebbe dunque relativo a come gli studenti comprendano, interpretino, analizzino e discutano il feedback e, in ultima analisi, su come essi lo applichino al loro compito e alle loro conoscenze. Prospettive recenti, in effetti, confermano che è fondamentale che gli studenti siano attivi nel processo di apprendimento per decodificare e interpretare il feedback, confrontarlo con il proprio lavoro, identificare eventuali aree di criticità e modalità per migliorarle, e riflettere su questo processo al fine di utilizzare il feedback in modo sempre più efficace (Carless & Boud, 2018). In tal senso, Carless (2019) ritiene che il feedback migliore sia quello che ha una prospettiva a lungo termine, generando pensiero, riflessione e azione differita: uno degli obiettivi del feedback sarebbe proprio quello di spingere gli studenti ad adottare nuove prospettive a lungo termine, modalità d'azione diversificate e una maggiore consapevolezza. Tuttavia, implementare tali processi in classi di grandi dimensioni può rivelarsi problematico (Raffaghelli et al., 2018; Ranieri et al., 2018), in quanto la numerosità delle persone coinvolte rappresenta un ostacolo, che induce frequentemente i docenti universitari ad utilizzare le lezioni frontali tradizionali come unica modalità didattica praticabile (Deslauriers et al., 2019),

con la conseguente esclusione di pratiche di feedback in itinere e di feedback fra pari.

In questo scenario, è evidente che l'introduzione, nei contesti educativi, di dispositivi digitali offre notevoli opportunità per creare ambienti ricchi di risorse, interattività, dialogo costruttivo e condivisione delle conoscenze (Grion & Cesareni, 2016), anche con classi numerose (Bozzi et al., 2021). Sistemi automatizzati di feedback, per esempio, consentirebbero ai docenti di alleggerire il loro carico di lavoro per la valutazione e il feedback in itinere e, soprattutto, di mettere in atto una valutazione personalizzata in tempo reale, fornendo agli studenti commenti rapidi, dettagliati e implementabili (Pauli & Ferrell, 2020).

In effetti, le forme di feedback automatizzato spaziano dai quiz online con feedback strutturato a strumenti raffinati come i sistemi di allerta precoce (Bañeres et al., 2020; Rodríguez et al., 2022) che supportano la consapevolezza degli studenti sulla loro auto-organizzazione (Molenaar, 2022). Si tratta di strumenti e modalità di erogazione diversificate: pannelli visivi e dashboard che informano gli studenti sui loro progressi e indirizzano la loro attenzione sul processo pedagogico (Yoo et al., 2015); mappatura semantica e analisi del testo che forniscono informazioni sulla qualità degli elaborati o delle partecipazioni ai forum (Lancho et al., 2018; Ke & Ng, 2019); sistemi adattivi che suggeriscono percorsi di apprendimento personalizzati (Brusilovsky & Peylo, 2003; Nguyen et al., 2018); sistemi di tutoring intelligenti (Crow et al., 2018) con volti più o meno umani che assistono gli studenti nei loro dubbi o suggeriscono attività per approfondire la loro esperienza di apprendimento (Grion et al., 2021).

In ogni caso, l'integrazione adeguata delle tecnologie, e in particolare l'accettazione delle tecnologie avanzate, come nel caso dei sistemi basati sull'Intelligenza Artificiale (IA), richiede un processo sempre più intenzionale e consapevole di conoscenza e interazione con gli stessi agenti automatizzati (Rodríguez et al., 2024; Foschi et al., 2024). Questo aspetto è particolarmente rilevante per i docenti universitari che si confrontano con i nuovi sistemi basati sui dati: devono ripensare i loro approcci pedagogici e didattici e intraprendere un processo di apprendimento progressivo relativo agli strumenti tecnologici (Raffaghelli, 2024). Oltre a ciò, le esperienze di fallimento, gli impatti poco chiari e persino dannosi, segnalati anche nella letteratura critica (Facer & Selwyn, 2021; Raffaghelli et al., 2025), richiedono una riflessione pedagogica attenta e l'incorporazione progressiva di tali sistemi, con una attenzione particolare alle esperienze di docenti e studenti.

Nonostante tali elementi critici, è generalmente condiviso il fatto che gli agenti e gli ambienti digitali presentino un enorme potenziale, poiché supportano la personalizzazione e la diversificazione delle esperienze di apprendimento,

favorendo la motivazione e il coinvolgimento e contribuendo all'efficacia dell'apprendimento (Bonaiuti & Dipace, 2021). Tuttavia, nel contesto nazionale la recente letteratura di settore (Doria & Picasso, 2024) evidenzia un limitato utilizzo in ambito educativo, di pratiche di feedback potenziate dall'uso della tecnologia, nonostante il crescente interesse, a livello internazionale, allo sviluppo di competenze pedagogiche valutative potenziate dai sistemi digitali (Redecker & Punie, 2017).

In questo contesto, il Progetto di Rilevante Interesse Nazionale (PRIN) "AI&F"[1], coinvolgente le Università di Padova (UniPD), Macerata (UniMC) e Bari Aldo Moro (UniBA), si propone di sviluppare una metodologia per l'utilizzo di un framework di Machine Learning (ML) open-source, finalizzato a supportare i docenti nel fornire feedback di alta qualità a gruppi numerosi di studenti, generando percorsi interattivi e trasformativi in una logica ecosistemica.

Il presente contributo riguarda la prima parte del progetto, che ha avuto come obiettivo l'analisi delle pratiche di feedback attualmente in uso nell'istruzione superiore. Per rispondere a tale obiettivo si è utilizzato un disegno esplorativo finalizzato a indagare le pratiche di feedback da parte dei docenti – con particolare attenzione ai sistemi di feedback automatizzato – e le percezioni e le esperienze d'uso degli studenti universitari rispetto al feedback, sia umano che automatizzato.

2. La ricerca

Per rispondere all'obiettivo di ricerca delineato, sono stati previsti due nuclei di ricerca (Figura 1) finalizzati ad analizzare, nelle tre Università coinvolte nel PRIN:

1) le pratiche di feedback e di feedback automatizzato messe in atto dai docenti nei loro insegnamenti;
2) le percezioni e le esperienze d'uso degli studenti universitari rispetto all'utilizzo del feedback e del feedback automatizzato.

1 Bando PRIN 2022 progetto codice 2022ZMYYTH titolo "Artificial intelligence & feedback for effective learning (AI&F)" finanziato dall'Unione Europea – NextGenerationEU. PNRR – Missione 4: Istruzione e ricerca, Componente C2: "Dalla ricerca all'impresa", Investimento 1.1 "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)"

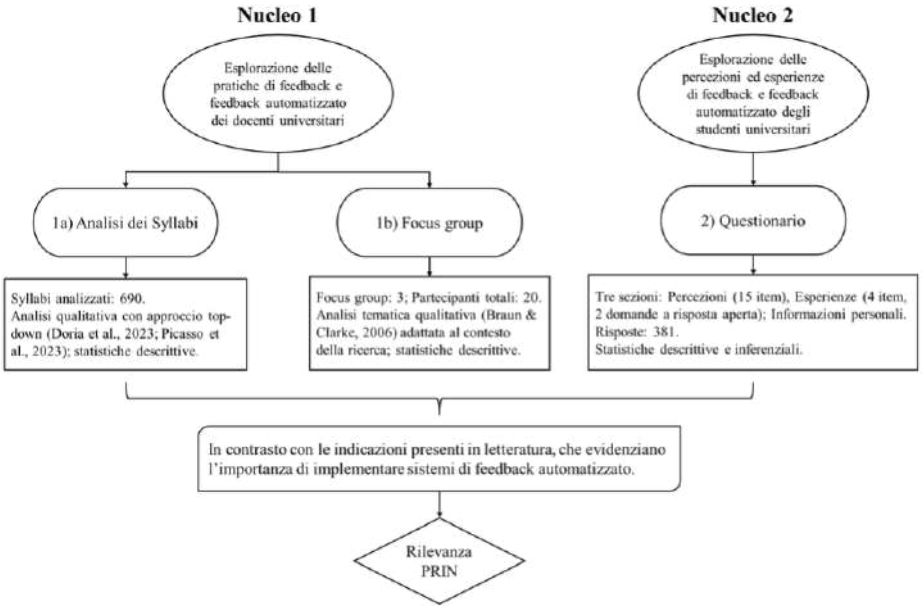


Fig. 1 – Flusso di lavoro del WP1

2.1 Obiettivo e domande di ricerca

L'obiettivo generale del presente lavoro è stato quello di indagare e analizzare, nel contesto delle tre Università coinvolte nel PRIN "AI&F", le concezioni e le pratiche di feedback di docenti delle tre università coinvolte e le concettualizzazioni e le esperienze di studenti delle stesse università, allo scopo di ottenere un quadro della situazione contestuale nella quale sarebbero state conseguentemente attuate le successive fasi della ricerca PRIN.

In relazione al primo nucleo di ricerca, le domande di ricerca a cui si è voluto dare risposta sono state le seguenti:

– RQ1. Ci sono – e quali sono – le pratiche di feedback e feedback automatizzato dichiarate dai docenti nei Syllabi dei loro insegnamenti?
– RQ2. Come viene definito il feedback e attraverso quali modalità e tipologie viene utilizzato dai docenti, incluso il feedback automatizzato con i relativi vantaggi e criticità?

In relazione al secondo nucleo di ricerca (2) si è proceduto con la costruzione e somministrazione di un questionario agli studenti, volto a indagare le loro percezioni circa l'utilità di specifiche situazioni e strumenti di feedback e di feedback automatizzato e le loro esperienze in merito. La domanda di ricerca a cui si è voluto dare risposta è stata la seguente:

– RQ3. Quali percezioni ed esperienze di feedback sperimentano le e gli studenti nelle tre università?

2.2 Metodo

2.2.1 Strumenti e procedure

Primo nucleo di ricerca

In relazione al primo nucleo di ricerca, sono state dapprima analizzate le pratiche di feedback dichiarate nei Syllabi (1a in Fig. 1) e successivamente si è approfondito come il feedback, incluso quello automatizzato, viene definito e utilizzato dai docenti, attraverso la conduzione di tre focus group (1b in Fig. 1).

Sono stati analizzati i Syllabi di un campione rappresentativo di docenti appartenenti alle tre università partner (UniPD: n=322; UniMC: n=150; UniBA: n=290), selezionato stratificando l'intera popolazione docente per i quattordici settori scientifico-disciplinari. I Syllabi completi analizzabili sono risultati 690. Su questi è stata condotta un'analisi qualitativa con approccio top-down (Ding, 2023), i cui risultati sono stati poi elaborati mediante analisi statistiche descrittive.

Per approfondire le pratiche di feedback adottate dai docenti nelle situazioni d'aula (1b in Fig. 1), sono stati condotti tre focus group in tre diversi dipartimenti, uno per ciascun Ateneo, con un campione di 20 docenti (6-8 per ateneo). I focus group hanno esplorato i significati assegnati al feedback (definizione, modalità di utilizzo, tipologie utilizzate) e al feedback automatizzato (utilizzo, vantaggi, criticità). Le trascrizioni delle interazioni discorsive sono state analizzate qualitativamente adattando allo specifico contesto della ricerca i passaggi procedurali di analisi tematica indicati da Braun e Clarke (2006), con successiva quantificazione delle occorrenze per ciascuna categoria.

Secondo nucleo di ricerca

Per indagare le percezioni e le esperienze degli studenti relativamente alle pratiche di feedback, è stato progettato un questionario sulla base della letteratura sul tema, composto da tre sezioni. Le tre sezioni principali sono state precedute

da una sezione introduttiva in cui sono stati presentati la ricerca e il questionario, fornite le definizioni di feedback e feedback automatizzato e illustrati gli aspetti relativi al consenso informato, inclusi privacy, norme etiche della ricerca e trattamento dei dati. Nel dettaglio, il questionario era composto come di seguito (alcune sezioni sono riportate anche in Appendice 1).

Sezione 1 – Situazioni. La sezione "Situazioni" era composta da quindici item e mirava ad analizzare le percezioni degli studenti circa l'utilità di diverse situazioni (S) e strumenti (T) di feedback. Le situazioni erano riferite a tre delle fonti primarie di feedback identificate da Panadero e Lipnevich (2022) e Grion e colleghe (2024): docente, pari, computer. Gli strumenti, invece, sono stati scelti tra quelli più utilizzati nel contesto italiano, i.e. ChatGPT, Grammarly, Compilatio Studium o Turnitin. Nello specifico, tre item erano relativi a *situazioni di feedback* da parte del docente, quattro a *situazioni di feedback tra pari*, cinque a *situazioni di feedback automatizzato* e tre a *strumenti di feedback automatizzato* (si veda Tab. 8). Per le diverse situazioni la scala di risposta era di tipo Likert da 1 (Completamente in disaccordo) a 7 (Completamente d'accordo), e per i tre strumenti la scala andava da 1 (Per nulla) a 7 (Del tutto).

Sezione 2 – Esperienze. La sezione "Esperienze" si componeva di 4 item con quattro opzioni di risposta (Mai, Qualche volta, Spesso, Sempre) e mirava a indagare le esperienze concrete degli studenti con diverse tipologie di feedback.

Sezione 3 – Anagrafica. L'ultima sezione raccoglieva informazioni personali, tra cui l'età, il genere, il livello di studio, il corso di laurea.

2.2.2 Campione e analisi dati

Primo nucleo di ricerca

Al fine di indagare le pratiche di feedback e feedback automatizzato messe in atto dai docenti (RQ1) sono stati considerati come oggetto di analisi i Syllabi. Non essendo reperibile un elenco nazionale di Syllabi, per formare il campione di analisi si è scelto di fare riferimento ai docenti – e ai loro rispettivi Syllabi – facenti parte delle università partner del progetto. È stato pertanto costruito un campione rappresentativo dei docenti appartenenti alle tre università partner del progetto PRIN (Cfr. Tab.1).

Università degli Studi di Padova	**Università degli Studi di Macerata**	**Università degli Studi di Bari**
Campione rappresentativo di Syllabi corrispondenti a 322 docenti	Campione rappresentativo di Syllabi corrispondenti a 150 docenti	Campione rappresentativo di Syllabi corrispondenti a 290 docenti

Tab. 1 – Campione dei docenti selezionato per ciascuna Università partner del progetto

Il campione è stato selezionato mediante stratificazione dell'intera popolazione docente appartenente alle Università partner (UniPD; UniMC; UniBA) in sottopopolazioni, quali i settori scientifico-disciplinari di appartenenza dei singoli docenti. È stato così individuato un campione rappresentativo di docenti per ciascuna universitàcon un errore statistico del 5% e un livello di confidenza del 95%.

Il campione totale è stato così composto come riportato in Tab. 2.

Area scientifico-disciplinare	Numerosità campione	% campione
A1. Scienze matematiche e informatiche	37	4.86%
A2. Scienze fisiche	28	3.67%
A3. Scienze chimiche	38	4.99%
A4. Scienze della terra	21	2.76%
A5. Scienze biologiche	60	7.87%
A6. Scienze mediche	98	12.86%
A7. Scienze agrarie e veterinarie	65	8.53%
A8. Ingegneria civile e Architettura	14	1.84%
A9. Ingegneria industriale e dell'informazione	40	5.25%
A10. Scienze dell'antichità, filologico-letterarie e storico-artistiche	80	10.50%
A11. Scienze storiche, filosofiche, pedagogiche e psicologiche	91	11.94%
A12. Scienze giuridiche	98	12,86%
A13. Scienze economiche e statistiche	68	8.92%
A14. Scienze politiche e sociali	24	3.15%
Totale	762	100%

Tab. 2 – Campione dei docenti per area scientifico-disciplinare

Successivamente, sono stati individuati 762 Syllabi, selezionando casualmente uno per ciascun docente facente parte del campione. I Syllabi, scaricati dai siti di ciascun Ateneo di riferimento, sono stati esaminati per verificare che vi fossero compilate le sezioni che avrebbero potuto includere elementi riguardanti le pratiche valutative utilizzate dai docenti, ossia: a) i risultati di apprendimento attesi; b) le pratiche didattiche; c) le modalità di valutazione.

Dai Syllabi-campione sono stati esclusi 72 Syllabi perché mancanti di informazioni (una o più sezioni d'interesse per l'analisi non erano compilate). I Syllabi analizzabili sono risultati così 690. Su questi ultimi è stata compiuta inizialmente un'analisi qualitativa, i cui risultati sono stati successivamente trattati con analisi statistiche descrittive.

L'analisi qualitativa del contenuto, condotta con approccio top-down basato sui framework teorici proposti da Doria et al. (2023) e Picasso et al. (2023), è stata svolta mediante il software di analisi testuale ATLAS.ti22. Due giudici indipendenti hanno codificato il 10% dei documenti per raggiungere una "sensibilità comune" nel processo di codifica, con un accordo del 95%. Successivamente, il restante corpus è stato suddiviso tra i due giudici, che hanno proseguito l'analisi in maniera indipendente, applicando il medesimo protocollo precedentemente descritto. Attraverso questa procedura, sono stati individuati nei testi i codici relativi alle pratiche valutative, che hanno permesso di classificare gli approcci valutativi (Guskey, 2019; Lipnevich et al., 2021) a cui i soggetti fanno riferimento per l'assegnazione del voto finale:

– *valutazione di prodotto*: misura ciò che gli studenti sono in grado di fare in un momento specifico, privilegiando una valutazione sommativa volta a certificare le competenze acquisite al termine del percorso formativo. I docenti che utilizzano questo approccio assegnano i voti principalmente in base ai punteggi ottenuti attraverso esami finali o prodotti conclusivi;
– *valutazione di processo*: considera non solo i risultati finali degli studenti, ma anche il percorso attraverso il quale acquisiscono competenze e conoscenze. I docenti che adottano questo approccio valutano sia il prodotto finale sia il processo di apprendimento, utilizzando strumenti come quiz autovalutativi, prove intermedie, attività individuali o di gruppo che incentivano la partecipazione degli studenti;
– *valutazione di progresso*: misura l'apprendimento in termini di avanzamento, valutando il miglioramento degli studenti da uno stato iniziale a uno finale. Questo approccio si concentra sul "guadagno d'apprendimento" e sulla crescita educativa, quantificando in modo concreto l'incremento del livello di competenze e conoscenze raggiunto dagli studenti grazie alle esperienze formative.

In relazione alla seconda domanda di ricerca (RQ2), sono stati organizzati tre focus group nelle seguenti aree scientifico-disciplinari previste dal progetto PRIN: *Humanities* presso l'Università di Padova, *Social Sciences* presso l'università di Macerata, e *Hard Sciences* presso l'Università di Bari. Il campione di convenienza, composto da 20 docenti, è stato selezionato in base alle disponibilità dei docenti all'interno dei dipartimenti coinvolti, con l'obiettivo di garantire una distribuzione equilibrata tra le tre università. In particolare, ogni focus group ha visto la partecipazione di 6-8 docenti per ciascun ateneo.

I focus group sono stati svolti durante il mese di marzo 2024, utilizzando la

piattaforma Zoom al fine di agevolare la raccolta dati e la loro trascrizione. Ogni incontro ha avuto una durata di circa 90 minuti, permettendo di approfondire le questioni trattate e di raccogliere informazioni dettagliate sui temi indagati (Tab. 3). Le discussioni sono state guidate da un facilitatore esperto, con l'obiettivo di promuovere un dialogo fluido e costruttivo tra i partecipanti. Gli incontri si sono sviluppati attorno a domande aperte, progettate per esplorare le percezioni dei docenti sulle pratiche di feedback adottate, le loro esperienze personali e le eventuali sfide incontrate. Le macrocategorie e le categorie analizzate sono presentate di seguito.

Macrocategorie	**Categorie**	**Definizione**
Feedback	Definizione di feedback	Concezione della natura stessa del feedback secondo i docenti nell'ambito dell'insegnamento universitario.
	Modalità di utilizzo del feedback	Analisi delle diverse modalità attraverso cui il feedback viene fornito e gestito nell'ambito dell'insegnamento universitario.
	Tipologie di feedback utilizzate	Rilevazione delle tipologie di feedback utilizzate dagli insegnanti e dagli studenti.
Feedback automatizzato	Uso del feedback automatizzato	Esplorazione dell'utilizzo di sistemi automatizzati per la generazione e la distribuzione del feedback dei docenti universitari.
	Vantaggi feedback automatizzato	Analisi dei benefici derivanti dall'utilizzo di sistemi automatizzati per la gestione del feedback.
	Svantaggi/Criticità feedback automatizzato	Analisi delle criticità e le limitazioni associate all'utilizzo di sistemi automatizzati per la gestione del feedback.

Tab. 3 – Macrocategorie e relative categorie di indagine dei focus group

Le trascrizioni dei focus group sono state analizzate mediante il software di analisi ATLAS.ti22. Una volta trascritto il testo, il processo di analisi è stato condotto seguendo i passaggi raccomandati da Braun e Clarke (2006), adattati al contesto della ricerca, come segue.

Lettura iniziale: in questa fase preliminare, il testo è stato letto per consentire una familiarizzazione con i contenuti e cogliere il tono generale delle discussioni. Questo ha permesso di identificare spunti iniziali e comprendere le dinamiche delle percezioni dei docenti riguardo al feedback.

Seconda lettura e codifica iniziale: successivamente, è stata effettuata una se-

conda lettura approfondita, avviando il processo di codifica. In questa fase, i codici sono stati assegnati alle parti rilevanti dei discorsi, senza utilizzare categorie prestabilite, permettendo così ai temi di emergere direttamente dai dati.

Integrazione delle categorie: durante la terza fase, una lettura mirata è stata condotta per integrare i codici emersi nelle categorie e macrocategorie previste dalla struttura del focus group. Questo passaggio ha assicurato che le risposte fossero coerentemente classificate all'interno dei temi principali.

Risoluzione dei problemi di codifica: la quarta lettura ha affrontato eventuali problemi di codifica, con particolare attenzione alla coerenza interna. È stata un'opportunità per correggere discrepanze e assicurare che ogni codice fosse coerente con le categorie stabilite.

Riflessione condivisa sui temi emersi: nella quinta fase, vi è stata una riflessione sui temi emersi durante il processo di codifica. Questa riflessione qualitativa ha consentito di interpretare e contestualizzare i dati raccolti, valorizzando le percezioni dei docenti in relazione alle pratiche di feedback.

Quantificazione: infine, nell'ultima fase, sono state quantificate le categorie emerse, combinando elementi qualitativi e quantitativi per ottenere una visione più completa. Questo processo è stato supportato dall'uso del software ATLAS.ti22, che ha facilitato l'analisi e la rappresentazione dei dati.

Secondo nucleo di ricerca

I questionari raccolti sono stati 381, suddivisi in 119 compilati da studenti/sse universitari attualmente frequentanti un corso di laurea e 262 da studenti/sse laureati/e (tra il 2010 e il 2024[2]) attualmente frequentanti percorsi formativi di abilitazione all'insegnamento (Cfr. Tab 4).

Categoria	Sottocategorie	% sul totale del campione
Fasce d'età	Meno di 23 anni	13.9
	23-29 anni	33.6
	30-35 anni	25.5
	36-40 anni	15.5
	41-45 anni	6
	Più di 45 anni	5.5
Genere	Femminile	70.1
	Maschile	27.8
	Preferisco non esplicitarlo	2.1

2 In particolare, il 12.6% (48) dei rispondenti si è laureato tra il 2010 e il 2014, il 19.9% (76) tra il 2015 e il 2019 e il 36.2% (138) tra il 2020 e il 2024.

Categoria	Sottocategorie	% sul totale del campione
Studenti	Laurea Triennale	19.2
	Laurea Magistrale a Ciclo Unico	4.5
	Laurea Magistrale	7.6
Laureati	Laurea Magistrale	58.5
	Dottorato di ricerca	3.4
	Laurea Magistrale a Ciclo Unico	3.4
	Doppia laurea (doppia LM o LM e CU)	.8
	Laurea Triennale	2.1
	Laurea del Vecchio Ordinamento	.5
Classi di laurea[3]	Filologia Moderna (LM-14)	10
	Beni Culturali (L-1)	8.4
	Scienze della Formazione Primaria (LM-85 bis)	7.6
	Scienze e Tecniche Psicologiche (L-24)	7.3
	Matematica (LM-40)	5
	Storia (L-42)	3.7
	Lingue e Letterature Europee e Americane (LM-37)	3.4
	Scienze e Tecnologie Agrarie (LM-69)	3.1
	Scienze dello Spettacolo e della Produzione Multimediale (LM-65)	2.4
	Linguistica (LM-39)	2.4
	Ingegneria Civile (LM-23)	2.1
	Biologia (LM-6)	2.1
	Scienze e Tecniche delle Attività Motorie Preventive e Adattate (LM-67)	1.8
	Scienze Statistiche (LM-82)	1.6
	Scienze Chimiche (LM-54)	1.6

Tab. 4 – Caratteristiche del campione

I dati raccolti sono stati analizzati utilizzando sia statistiche descrittive sia test inferenziali.

Per valutare le esperienze degli studenti con diverse tipologie di feedback, sono state calcolate le frequenze delle risposte fornite per ciascuna delle quattro categorie ordinali della scala. Per analizzare se esistessero differenze nella frequenza con cui gli studenti avevano sperimentato le quattro tipologie di feedback, sono stati adottati test non parametrici per misure ripetute (Friedman seguito da post-hoc con uso di Wilcoxon e correzione di Bonferroni).

3 Sono riportate unicamente le classi di laurea indicate da 6 o più studenti/laureati.

Per l'analisi delle situazioni e degli strumenti di feedback, sono stati calcolati indici di tendenza centrale e dispersione per descrivere la distribuzione delle risposte. Successivamente è stato applicato il t-test per campioni singoli per confrontare le medie osservate con un valore teorico di riferimento (i.e., il punto medio della scala, 4). Infine, la dimensione dell'effetto è stata calcolata e riportata (d di Cohen).

In aggiunta, gli item relativi alle situazioni delle tre fonti di feedback sono stati aggregati al fine di formare tre scale: Docente, Pari e Computer. Tale aggregazione è stata resa possibile dalla caratterizzazione teorica dei costrutti. La coerenza interna delle tre scale costruite è stata analizzata adottando gli indici alpha di Cronbach (α) e omega di McDonald (ω). Per ciascuna scala, analogamente a quanto fatto per i singoli item, è stato verificato se la media si discostasse significativamente dal punto medio della scala di misura mediante un t-test per campioni singoli, accompagnato dalla stima della dimensione dell'effetto. Successivamente, al fine di confrontare le diverse fonti di feedback (i.e., Docente, Pari e Computer), si è eseguita un'analisi della varianza (ANOVA) per misure ripetute, con relativa analisi post-hoc utilizzando il metodo di correzione di Bonferroni e calcolo della dimensione dell'effetto (eta quadrato).

In generale, il livello di significatività (p) è stato fissato a .05 per tutte le analisi inferenziali. Tutte le elaborazioni sono state condotte mediante software statistico dedicato (Jamovi).

3. Risultati

RQ1: Ci sono - e quali sono – le pratiche di feedback e feedback automatizzato dichiarate dai docenti nei Syllabi dei loro insegnamenti?

Il campione di docenti (*M*=56.91%; *F*=43.09%), da cui sono stati selezionati casualmente i Syllabi, è composto principalmente da professori associati (60.45%), seguiti da professori ordinari (30.39%) e da ricercatori (9.16%).

In merito alla tipologia di approccio al voto finale che i docenti universitari utilizzano, i dati confermano che, naturalmente, tutti i docenti (*n*=690, 100% dei docenti) dichiarano di utilizzare una valutazione di *prodotto*, ossia volta a verificare gli esiti degli apprendimenti degli studenti, ma solamente il 25.8% di loro (*n*=178) dichiara di mettere in atto anche una valutazione di *processo*. In nessun caso si è riscontrata una dichiarazione di utilizzo di una valutazione di *progresso*.

In riferimento alle *pratiche di valutazione messe in atto nell'ambito dei loro insegnamenti,* i dati codificati rilevano che le pratiche di valutazione maggiormente citate dai docenti nei loro insegnamenti sono gli esami finali orali (39.81% di occorrenze sul totale) e/o gli esami scritti finali (24.69% di occorrenze sul totale).

Le altre pratiche valutative dichiarate dai docenti sono utilizzate con minore frequenza, come si evince dalla Tab. 5.

Pratiche di valutazione		Fq delle occorrenze	% delle occorrenze
Valutazione di *prodotto*	Esame orale finale individuale	479	39.81%
	Esame scritto finale individuale	297	24.69%
	Esame pratico finale individuale	31	2.58%
	Esame parziale *(es. prova sommativa in itinere)*	85	7.07%
Valutazione di *processo*	Attività individuali in itinere (report, project work, relazione orale)	119	9.89%
	Attività di gruppo in itinere (report, project work, relazione orale)	48	3.99%
	Altre pratiche di valutazione autentica (prove *real life*, ecc)	18	1.50%
	Partecipazione e frequenza	37	3.08%
	Feedback	89	7.40%
Valutazione di *progresso*	Valutazione diagnostica individuale	0	0%
Totale occorrenze		1203	100%

Tab. 5 – Frequenze e % delle occorrenze relative alle pratiche di valutazione citate dai docenti, classificate in base al criterio di prodotto, processo e progresso

In relazione alle *pratiche di feedback,* i dati evidenziano che il 9.71% (*n*=67) dei docenti dichiara di utilizzare nei propri insegnamenti pratiche legate all'uso di feedback. In particolare, il feedback è utilizzato come strumento per favorire l'autovalutazione (48 occorrenze sul totale delle 89 come da Tab. 5) e feedback dato dal docente (24 occorrenze). Solamente un docente (.14% dei docenti, .08% delle occorrenze), invece, dichiara di utilizzare il feedback automatizzato (nell'accezione di *Computer Based Assessment practices* (CBA)[4] (Sim et al., 2004; Tonelli et al., 2018) (Tab 6).

4 Per computer based assessment (CBA) si intende una forma di valutazione in cui prove e compiti sono progettati, erogati e spesso corretti automaticamente tramite tecnologie digitali, con il computer che registra le risposte degli studenti, le valuta e può fornire feedback immediato, sostituendo o integrando i tradizionali test cartacei (Shute & Rahimi, 2017)

Pratiche di Feedback	Fq	%
Autovalutazione	48	3.99%
Feedback docente	24	1.99%
Feedback tra pari	14	1.16%
Revisione tra pari	2	.17%
Feedback CBA	1	.08%

Tab. 6 – Frequenze e % delle pratiche di Feedback citate dai docenti

RQ2: Come viene definito il feedback e attraverso quali modalità e tipologie viene utilizzato dai docenti, incluso il feedback automatizzato con i relativi vantaggi e criticità?

L'analisi delle interazioni discorsive dei focus group ha permesso di evidenziare tre macrocategorie di significati emergenti: 1) le definizioni di feedback; 2) le modalità di utilizzo; 3) le tipologie di feedback utilizzate.

Riguardo alla prima macrocategoria, emerge che il feedback viene prevalentemente definito come un processo di interazione tra docente e studente, come evidenziato dal codice più rappresentativo "Interazione studente-docente" (Tab. 7).

In termini di modalità di utilizzo, la "Discussione in aula" risulta la pratica più frequente per attivare pratiche di feedback (7.80%), mentre, rispetto alle tipologie, vi è una prevalenza del "Feedback studente-docente" (26.83%), seguito dal "Feedback docente-studente" (14.63%).

Macrocategoria: Feedback				
Categorie	**Codici**	**Quotazioni**	**%**	**Citazione**
Definizione di feedback	Interazione studente-docente	24	11.71%	"*Lezione partecipata, ossia, ogni tot comunque invito sempre a fermarmi a domandarmi cose se non è stato compreso quello che ho spiegato"* 2:2 ¶ 8 in Padova
	Approfondimento e riflessione	5	2.44%	*"Gli studenti prima o dopo lezione vengono a chiedermi cose per cui io rispondo puntualmente [... oppure] rispondo a tutti quanti [... o] rispondo via mail oppure propongo un incontro su zoom"* 2:4 ¶ 12-15 in Padova
	Ricaduta sull'apprendimento	3	1.46%	*"Feedback è quando [...] lo studente riesce a fare collegamenti tra un argomento e l'altro senza [...]"* 1:115 ¶ 20 in Macerata
	Commento	2	.98%	*"Il feedback che do rispetto a ogni singolo passaggio del prodotto che devono costruire e quindi la revisione. [... Una] descrizione, commento appunto di quello che hanno fatto."* 2:9 ¶ 19-23 in Padova
	Valutazione	1	.49%	*"Attraverso insegnamenti con prove parziali, o con più prove, noi riusciamo a dare un feedback in qualche modo a una serie di cose, che però si riducono al voto su quella parte."* 2:53 ¶ 219 in Padova
	Patto Formativo	1	.49%	"*Accordo su quello che si vuole trasmettere*" 1:103 ¶ 21- in Macerata
Modalità di utilizzo del feedback	Discussione in aula	16	7.80%	*"Se ne discuteva in Aula insieme nel caso in cui non fosse un lavoro facilissimo, anche perché poi io volevo anche tener conto di questo lavoro per l'esame. […] Però questa abitudine diciamo alla discussione in classe, insomma, è rimasta chiaramente. "* 2:41 ¶ 164 in Padova
	Azione di miglioramento	11	5.37%	*"Feedback per sapere se [...] hanno recepito bene poi se non l'hanno fatto dedico un'altra parte della lezione per approfondire quella parte."* 1:126 ¶ 129 in Macerata
	Valutazione	1	.49%	*"Le risposte da parte loro nella partecipazione spesso [..] hanno pregiudizio del giudizio, della valutazione in itinere, di tante cose che li condizionano"* 1:53 ¶ 185-189 in Macerata
	Patto Formativo	1	.49%	*"[Gli studenti] sono anche abbastanza sinceri e questo per me è molto utile per costruire il corso dell'anno successivo [... e] per la coprogettazione del corso successivo. In genere, all'inizio dell'insegnamento si prende il syllabus e si guarda insieme, [...] mi lascio un syllabus aperto."* 2:30 ¶ 113 in Padova

Tipologie di feedback utilizzate	Feedback Studente-Docente	55	26.83%	*"Anch'io stesso avrei bisogno di più feedback. [...] Ogni tanto, ho difficoltà a capire, diciamo, la base di partenza che ho."* 2:42 ¶ 166 in Padova
	Feedback Docente-Studente	30	14.63%	*"Cerco [...] di dare feedback continuo [...] li metto a lavorare in gruppi e giro tra i vari gruppi per dare il loro feedback, su come stanno lavorando e per dare loro una mia valutazione, un mio commento, diciamo su come stanno impostando il lavoro."*
	Feedback in ambiente digitale (posta elettronica; WooClap)	6	2.93%	*"L'utilizzo di Wooclap con le nuvole di parole [... chiedo] "Cosa vi aspettate da questa lezione? O meglio, secondo voi di cosa si va a parlare?" Cioè io vi do un titolo, che è una definizione, secondo voi cosa voglio andare?"* 2:47 ¶ 196 in Padova
	Feedback tra pari	1	.49%	*"L'idea della co-valutazione [...] la utilizzo quindi in funzione della preparazione alle prove che hanno anche magari una parte più pratica."* 2:18 ¶ 57-58 in Padova
	Exemplar	1	.49%	*"Faccio dei momenti in cui o danno un feedback agli esercizi degli anni precedenti che mostro loro o provano loro a fare qualcosa e dei feedback."* 2:18 ¶ 57-8 in Padova
	Voto	1	.49%	"*Certo, attraverso insegnamenti con prove parziali o con più prove noi riusciamo a dare un feedback in qualche modo una serie di cose, di pezzi di feedback che però si si riducono al voto su quella parte"* 2:53 ¶ 219 in Padova

Tab. 7 – Risultati dei focus group relativi alla macrocategoria "Feedback"

Con specifico riferimento alla macrocategoria "Feedback automatizzato" (Tab. 8), che rappresenta un focus di particolare interesse del PRIN, l'utilizzo appare più limitato e concentrato principalmente su strumenti istituzionali, in particolare Moodle (3.41%), impiegato soprattutto per finalità formative e di autovalutazione (1.95%), nonché per pratiche formative online (1.46%) e, in misura minore, per il feedback tra pari (.49%). Tra i vantaggi percepiti emerge la tempestività (2.43%) del feedback automatizzato, mentre le principali criticità riguardano l'incompletezza del feedback (2.93%) e una certa sfiducia nei sistemi automatizzati (2.93%).

Macrocategoria: Feedback automatizzato				
Categorie	**Codici**	**Quotazioni**	**%**	**Citazione**
Uso del feedback automatizzato	Moodle	7	3.41%	"*strumento feedback di Moodle.*" 2:14 ¶ 52 in Padova
	Nessun utilizzo	4	1.95%	"*Metodi bellissimi che utilizzano altri colleghi [...] appunto, non li ho usati.*" 1:60 ¶ 284 in Macerata
	Autovalutazione	4	1.95%	"*È uno strumento di autovalutazione, perché magari mi rimanda anche dove andare a cercare, appunto, se la risposta non era quella corretta.*" 1:100 ¶ 709 in Macerata
	Pratiche formative online	3	1.46%	"*Io faccio delle simulazioni della prova di esame [...] poi vengono discusse in classe, cioè ogni soprattutto le domande aperte, le esigenze delle fonti vengono discusse in classe.*" 2:44 ¶ 167 in Padova
	Nessuna conoscenza	2	.98%	"*E quindi, non ne so abbastanza da potermi esporre, esprimere.*" 1:109 ¶ 741 in Macerata
	Feedback tra pari	1	.49%	"*Io utilizzo la co-valutazione che è praticamente parte dell'esame, un esercizio di co-valutazione.*" 2:56 ¶ 237-239 in Padova
Vantaggi	Tempestività	5	2.43%	"*Essendo i dati in formato digitale, è facile anche trasformarli nel World cloud, cioè dare una visualizzazione anche agli studenti di quello che è venuto fuori delle loro risposte.*" 1:149 ¶ 490 in Macerata
	Supporto Docenti	3	1.46%	"*Aiuta molto il docente*" 1:96 ¶ 659 in Macerata
Svantaggi	Incompletezza	6	2.93%	"*Siccome c'è qualche studente, qualche studentessa che magari ha un discorso, usa parole, un discorso più complesso, più in profondità, mi piace la Carta e la penna, lascio qualche minuto in più.*" 2:38 ¶ 152 in Padova
	Sfiducia	6	2.93%	"*So che esistono sistemi ma mi preoccupano un po', come mi preoccupava all'inizio la traduzione automatica, nel senso che dipende tutto da come funziona il sistema che poi va a dare feedback, no? [...] Però ecco, mi piacerebbe capire come funzionano i sistemi [... perchè] c'è voluto un po' di tempo prima di metterli a punto*" 1:110 ¶ 741-745 in Macerata
	Necessità formative	4	1.95%	"*A me personalmente manca, cioè qualcuno che anche magari del mestiere che mi, sì un pochino, mi indirizzi, mi presenti delle varie opzioni*" 2:58 ¶ 246 in Padova
	Asettico	1	.49%	"*Soprattutto rendere le cose così asettiche.*" 2:60 ¶ 251 in Padova

Tab. 8 – Risultati dei focus group relativi alla macrocategoria "Feedback automatizzato"

RQ3. Quali percezioni ed esperienze di feedback sperimentano le e gli studenti nelle tre università?

Come mostrato in Tab. 9, i risultati non rispecchiano del tutto quanto mostrano i risultati delle analisi sul campione dei docenti. Il dato maggiormente evidente, e confermato da entrambe le popolazioni considerate, docenti e studenti, riguarda il fatto che le pratiche di feedback non risultano molto diffuse nei contesti formativi universitari. Va infatti rilevato che per tutti e quattro gli item relativi alle possibili esperienze di feedback nelle attività d'insegnamento/apprendimento, si registra una percentuale estremamente bassa di studenti che ha risposto di avere sperimentato "Sempre" tali situazioni. Le percentuali di coloro che hanno risposto "Mai" risultano superiori a quelle di chi ha risposto "Spesso", tranne nel caso di E1 (feedback dal docente), in cui le risposte "Spesso" risultano lievemente superiori alle risposte "Mai".

Con un certo interesse si osserva che per le esperienze E2 (Feedback tra pari) ed E3 (Auto-feedback), la maggioranza assoluta della popolazione intervistata ha risposto "Qualche volta", mentre per E4 (Feedback automatizzato) individuato con percentuali molto esigue nelle risposte dei docenti - è la maggioranza relativa (43% e 46.5%, rispettivamente). Questi ultimi risultati risultano più ottimistici sull'uso di pratiche di feedback nelle aule universitarie, in confronto a quanto osservato nel campione dei docenti, ma rimangono comunque valori molto bassi in relazione ad uno strumento - quale il feedback - di grande impatto sull'apprendimento.

		Mai		**Qualche volta**		**Spesso**		**Sempre**	
Item	Esperienza	Fq	%	Fq	%	Fq	%	Fq	%
E1	Feedback docente	100	26.2%	164	43%	103	27%	14	3.7%
E2	Feedback tra pari	120	31.5%	193	50.7%	64	16.8%	4	1%
E3	Auto-feedback	115	30.2%	202	53%	61	16%	3	.8%
E4	Feedback automatizzato	128	33.6%	177	46.5%	71	18.6%	5	1.3%

Tab. 9 – Risultati della sezione "Esperienze"

Per analizzare se esistessero differenze nella frequenza con cui gli studenti dichiarano di avere sperimentato le quattro tipologie di feedback, è stato applicato il test di Friedman sulle risposte trasformate in punteggi ordinali (1-4). Il test si

è rivelato significativo, $\chi^2(3) = 39.8$, $p < .001$. I successivi test post-hoc, condotti mediante confronti a coppie con test di Wilcoxon per campioni appaiati (ipotesi bidirezionale) e interpretati applicando la correzione di Bonferroni del livello di significatività ($\alpha_{Bonferroni} = .05/6 = .0083$), hanno mostrato l'assenza di differenze significative tra E2 ed E3, E2 ed E4 ed E3 ed E4 (tutti $p > \alpha_{Bonferroni}$), mentre sono emerse differenze statisticamente significative tra E1 (feedback docente) e ciascuna delle altre tre esperienze di feedback (tutti $p < \alpha$Bonferroni). In particolare, E1 (Me = 2; Q1 = 1, Q3 = 3) presenta una mediana superiore rispetto a E2 (Me = 2; Q1 = 1, Q3 = 2), E3 (Me = 2; Q1 = 1, Q3 = 2) ed E4 (Me = 2; Q1 = 1, Q3 = 2).

In definitiva, pur in un contesto complessivamente caratterizzato da un'esperienza limitata[5], i risultati evidenziano che il feedback del docente è quello sperimentato con una significativa maggiore frequenza dagli studenti.

Per quanto riguarda la sezione "situazioni e strumenti" di feedback, indagate nel questionario con gli item S1-S15, di seguito è proposta una sintesi dei risultati suddivisi per fonte come da quadro teorico assunto a fondamento dell'analisi (Panadero e Lipnevich, 2022).

Docente

Tutte e tre le situazioni proposte nel questionario, di feedback da parte del docente (commenti scritti; commenti orali e osservazioni generali), sono state percepite dagli studenti come decisamente utili per migliorare il proprio lavoro. Inoltre, gli item classificati in base all'utilità percepita, per ordine decrescente in relazione alla dimensione dell'effetto risultano: S1, S2, S3 (Tab.10). La situazione percepita come più utile è quindi quella relativa ai feedback scritti del docente, seguita dai suoi feedback orali e da quelli generali in aula.

In aggiunta, aggregando tutti gli item relativi alle situazioni di feedback docente per calcolare una scala complessiva, anche la media di quest'ultima risulta significativamente superiore al punto medio della scala di misura, con una dimensione dell'effetto gigante.

5 La limitata esperienza con le diverse tipologie di feedback non emerge soltanto dalle frequenze riportate in Tab. 9, è confermata anche dall'analisi inferenziale. Applicando il test di Wilcoxon per campioni singoli, la mediana delle risposte di ciascuno dei quattro item risulta infatti significativamente inferiore al punto mediano della scala di misura (i.e., $Me_0 = 2.5$, *p*s < .001).

					Percentili			Test t di Student			
Item	**Situazione di feedback**	**N**	**M**	**DS**	**25°**	**50° (Me)**	**75°**	**Statistica del Test**	**gdl**	**p**	**d**
S1	Feedback scritti del docente	381	6.08	1.02	6	6	7	40	380	<.001	2.05
S2	Feedback orali del docente	381	5.87	1.14	5	6	7	32	380	<.001	1.64
S3	Feedback generali in aula del docente	381	5.83	1.16	5	6	7	30.8	380	<.001	1.58
	Docente $\alpha = .74$; $\omega = .77$	381	5.93	.899	5.33	6	6.67	41.8	380	<.001	2.14

Tab. 10 – Risultati della sezione "Situazioni" – Docente
Test t di Student. H1 è di tipo bidirezionale, i.e. $M \neq 4$. Il livello di significatività (p) fissato per il rifiuto di H0 è .05

Pari

Tutte e quattro le situazioni di feedback da parte dei pari sono state percepite come decisamente utili per migliorare il proprio lavoro. Infatti, la media delle risposte di ciascuno dei quattro item è risultata significativamente superiore al punto medio della scala (*p*s < .001; si veda Tab. 11) con dimensioni dell'effetto *grandi* per S5, S6 e S7 e *gigante* per S4. Inoltre, gli item classificati in ordine decrescente in relazione a queste ultime risultano: S4, S6, S5, S7. La situazione percepita come più utile è quindi quella relativa ai feedback scritti dei pari, seguita dai loro feedback generali in aula, da quelli orali e dai feedback informali fuori dall'aula.

Tuttavia, nonostante l'utilità percepita, l'esperienza effettiva degli studenti con questa tipologia di feedback in contesti universitari è limitata. In particolare, come mostrato in relazione alle *Esperienze* relative ai processi di feedback tra pari (si veda Tab. 11), la maggior parte degli studenti ha dichiarato di aver frequentato corsi in cui sono stati stimolati esplicitamente tali processi solo "Qualche volta" (50.7%) o addirittura "Mai" (31.5%), mentre solo una minoranza "Spesso" (16.8%) e solo quattro studenti hanno scelto "Sempre" (1%).

					Percentili			Test t di Student			
Item	**Situazione di feedback**	**N**	**M**	**DS**	**25°**	**50° (Me)**	**75°**	**Statistica del Test**	**gdl**	***p***	***d***
S4	Feedback scritti dei pari	381	5.32	1.24	5	5	6	20.7	380	<.001	1.06
S5	Feedback orali dei pari	381	5.16	1.33	4	5	6	17.1	380	<.001	.87
S6	Feedback generali in aula dei pari	381	5.22	1.34	4	5	6	17.8	380	<.001	.91
S7	Feedback dei pari fuori dall'aula	381	5.12	1.32	4	5	6	16.5	380	<.001	.85
	Pari $\alpha = .87$; $\omega = .87$	381	5.21	1.11	4.75	5.25	6	21.3	380	<.001	1.09

Tab. 11 – Risultati della sezione "Situazioni" – Pari

Computer

Tutte e cinque le situazioni di feedback automatizzato indicate sono state percepite dagli studenti come decisamente utili per migliorare il proprio lavoro e tutti gli strumenti di feedback automatizzato proposti nel questionario sono stati percepiti come decisamente in grado di aiutarli nel loro lavoro (Tab. 12). Infatti, la media delle risposte di ciascuno degli otto item è risultata significativamente superiore al punto medio della scala (ps < .001; si veda Tab.12), con dimensioni dell'effetto *giganti* per S14, *grandi* per S8, S13-T, S10 e S9, *moderate* per S15, S12-T e S11-T. La situazione percepita come più utile è quella relativa ai feedback generati da sistemi di analisi del testo (S14), seguita dai feedback generati dalle analitiche dei sistemi di gestione dell'apprendimento LMS (S8). Per quanto riguarda gli strumenti di feedback, invece, quelli percepiti come più utili sono i software antiplagio (S13-T), seguiti dai correttori automatizzati della grammatica (S12-T) e, infine, dai chatbot (S11-T).

Tuttavia, come nel caso del feedback tra pari, nonostante l'elevata utilità percepita, come precedentemente rilevato, l'esperienza effettiva degli studenti con questa tipologia di feedback in contesti universitari è assai limitata.

					Percentili			Test t di Student			
Item	**Situazione o strumento di feedback**	**N**	**M**	**DS**	**25°**	**50° (Me)**	**75°**	**Statistica del Test**	**gdl**	**p**	**d**
S11-T	Feedback da chatbot: ChatGPT	381	4.78	1.40	4	5	6	10.9	380	<.001	.56
S12-T	Feedback da correttori automatizzati della grammatica: Grammarly	381	5.02	1.43	4	5	6	13.9	380	<.001	.71
S13-T	Feedback da software antiplagio: Compilatio Studium o Turnitin	381	5.33	1.36	5	6	6	19	380	<.001	.97
S8	Feedback da analitiche LMS	381	5.25	1.29	4	5	6	18.9	380	<.001	.97
S9	Feedback automatizzati su piattaforma e-learning	381	5.18	1.34	4	5	6	17.2	380	<.001	.88
S10	Feedback da bacheche e grafiche di applicazioni	381	5.16	1.28	4	5	6	17.8	380	<.001	.91
S14	Feedback da sistemi di analisi del testo	381	5.34	1.24	5	6	6	21.1	380	<.001	1.08
S15	Feedback da sistemi di analisi dei forum	381	5.03	1.31	4	5	6	15.3	380	<.001	.79
	Computer $\alpha = .81$; $\omega = .81$	381	5.19	.97	4.6	5.4	5.8	23.9	380	<.001	1.23

Tab. 12 – Risultati della sezione "Situazioni" – Computer

Docente vs Pari vs Computer

Per analizzare se vi fossero differenze nei punteggi medi degli studenti alle tre scale (i.e., *Docente*, *Pari*, *Computer*[6]), è stata eseguita un'analisi della varianza per misure ripetute. Il test si è rivelato significativo, $F(1.9, 721.2^{7}) = 113$, $p < .001$, $\eta^2 = .106$[8]. I successivi test post-hoc, eseguiti con correzione Bonferroni del livello di significatività, hanno mostrato come non vi fossero differenze significative tra i punteggi medi delle scale *Pari* e *Computer* ($t(380) = .227$, $p_{Bonferroni} > .05$), mentre i punteggi medi della scala *Docente* risultavano significativamente superiori a entrambi quelli delle altre scale (rispettivamente: *Docente* vs *Pari*, $t(380) = 14.147$, $p_{Bonferroni} < .001$, $d = .725$; *Docente* vs *Computer*, $t(380) = 13.437$, $p_{Bonferroni} < .001$, $d = .688$[9]). In altre parole, confrontando le tre fonti di feedback (i.e., docente, pari, computer), è emerso che il feedback del docente risulta complessivamente percepito come più utile (M = 5.93, DS = .90) rispetto al feedback dei pari (M = 5.21, DS = 1.11) e al feedback generato dai sistemi digitali (M = 5.19, DS = .97).

Inoltre, considerando per ciascuna scala la dimensione dell'effetto associata allo scostamento rispetto al punto medio della scala, l'ordinamento decrescente risulta: *Docente* ($d = 2.14$), *Computer* ($d = 1.23$) e *Pari* ($d = 1.09$). Analogamente, esaminando i singoli item relativi alle situazioni di feedback e ordinandoli in base alla rispettiva dimensione dell'effetto, si ottiene che sia la scala Docente, sia le tre situazioni di feedback docente (S1-S3) occupano le prime posizioni, corrispondenti agli scostamenti più marcati dal punto medio e, quindi, ai livelli più elevati di utilità percepita.

4. Discussione e conclusioni

I risultati del presente studio offrono un quadro articolato sulle pratiche di feedback presenti nelle aule universitarie delle tre università coinvolte. Tale quadro è costruito attraverso le concettualizzazioni proprie dei docenti rispetto al feedback e alle percezioni e le esperienze di studenti universitari. Seppure i risultati

6 Per confrontare le tre fonti di feedback (i.e., docente, pari, computer), è stato utilizzato il punteggio aggregato dei diversi item relativi a ciascuna fonte, corrispondente alla scala complessiva a cui si è già fatto riferimento precedentemente.

7 Poiché l'assunzione di sfericità delle varianze non è stata soddisfatta (Test di sfericità di Mauchly $p < .05$), è stata utilizzata la correzione dei gradi di libertà di Greenhouse-Geisser.

8 Secondo le linee guida di Cohen (1988), tale valore corrisponde a un effetto *moderato*.

9 In entrambi i casi, secondo le linee guida di Cohen (1988), la differenza tra le medie ha una dimensione dell'effetto *moderata*.

abbiano permesso di descrivere in modo articolato il fenomeno indagato, essi vanno letti tenendo conto di alcune limitazioni.

Riguardo al campione di docenti, va rilevato che i focus group sono stati condotti con un numero esiguo di docenti e che il campione di studenti appartenenti alle tre università italiane era circoscritto, il che limita la generalizzabilità dei risultati. Inoltre, sebbene siano state analizzate diverse tipologie di feedback automatizzato, ulteriori ricerche potrebbero approfondire l'impatto specifico di ciascuna di queste, su differenti strategie di autoregolazione dell'apprendimento, nonché indagare l'efficacia di modelli didattici innovativi che prevedano un'integrazione più strutturata del feedback automatizzato con altre forme di valutazione formativa.

Nonostante tali limiti, si può ritenere di essere giunte ad individuare un quadro abbastanza ampio e articolato, a partire dal quale, da una parte si è potuto innestare il percorso di ricerca successivo con maggiore consapevolezza del contesto, dall'altra si offrono spunti di riflessione interessanti riguardo a come il feedback venga percepito e praticato in contesti accademici italiani rilevando potenzialità e criticità su cui agire per mettere in atto progressivi miglioramenti nelle pratiche didattiche future. Considerare attentamente le percezioni degli stakeholder rappresenta, d'altra parte, un modo per fare i conti con quelle *realtà percepite*, che rappresentano i curricoli nascosti di cui la ricerca educativa concorda nel rilevare il forte impatto sulle reali pratiche d'aula e sui risultati d'apprendimento perseguiti (Ergas, 2017; Sullivan, 2018).

Considerando più specificamente i risultati relativi al primo nucleo della ricerca, si possono proporre alcune considerazioni.

In primo luogo, l'analisi dei syllabi evidenzia una netta prevalenza di approcci valutativi centrati solo sul prodotto finale. Solo una minoranza di docenti (25.8%) fa riferimento a pratiche riconducibili alla valutazione di processo e nessun syllabus menziona esplicitamente approcci di valutazione di progresso. Questo dato appare coerente con quanto evidenziato nella letteratura internazionale sull'istruzione superiore, che sottolinea come le pratiche valutative universitarie continuino spesso a privilegiare forme sommative di certificazione degli apprendimenti piuttosto che processi di accompagnamento e regolazione dell'apprendimento lungo il percorso (Carless, 2015; Doria, 2025; Winstone et al., 2017).

All'interno di questo quadro, il feedback emerge come una pratica poco citata nei Syllabi: solo una minoranza di docenti (9.71%) dichiara esplicitamente l'utilizzo di pratiche di feedback e, tra queste, la forma più frequentemente espressa riguarda l'autovalutazione degli studenti. Ancora più marginale risulta il ricorso al feedback automatizzato, che compare in un solo caso. Questo risultato con-

ferma quanto già evidenziato da studi precedenti nel contesto italiano, che segnalano una limitata integrazione del feedback come pratica didattica strutturata nei dispositivi formali di progettazione didattica (Doria & Picasso, 2024).

Un elemento interessante che emerge dall'analisi dei focus group riguarda tuttavia la possibile discrepanza tra quanto dichiarato nei syllabi e quanto effettivamente praticato nelle aule universitarie. Nei focus group, infatti, i docenti tendono a definire il feedback prevalentemente come un processo dialogico di interazione con gli studenti, spesso situato nella discussione in aula o in momenti informali di confronto. Questa concezione appare coerente con le prospettive socio-costruttiviste che interpretano il feedback come processo dialogico e partecipativo (Carless & Boud, 2018; Nicol & MacFarlane-Dick, 2006). Tuttavia, tale interpretazione sembra tradursi in pratiche spesso poco strutturate e non sistematicamente integrate nella progettazione didattica formale.

Analogamente, l'uso di strumenti digitali per il feedback appare ancora limitato. Nei focus group, l'utilizzo di sistemi automatizzati si concentra principalmente su strumenti istituzionali, come le funzionalità di Moodle, spesso impiegate per attività di autovalutazione o simulazione delle prove d'esame. Inoltre, accanto ai vantaggi riconosciuti, in particolare la tempestività del feedback, emergono anche alcune criticità percepite dai docenti, tra cui il timore di incompletezza dei feedback generati automaticamente e una certa diffidenza nei confronti dei sistemi automatizzati. Queste percezioni riflettono le tensioni evidenziate nella letteratura recente sull'adozione delle tecnologie educative basate su dati e intelligenza artificiale, che richiedono non solo competenze tecniche ma anche una riflessione pedagogica approfondita sul loro ruolo nei processi di insegnamento e apprendimento (Raffaghelli, 2024; Facer & Selwyn, 2021).

Nel complesso, i risultati del primo nucleo di ricerca suggeriscono che, sebbene il feedback sia riconosciuto dai docenti come elemento rilevante nella relazione didattica, esso appare ancora poco formalizzato e raramente integrato in modo sistematico nelle pratiche valutative e nei dispositivi di progettazione dei corsi. Questo elemento ha rappresentato un aspetto cruciale considerato nelle successive fasi del progetto PRIN, che hanno mirato a sviluppare modelli pedagogici e strumenti tecnologici capaci di sostenere pratiche di feedback più strutturate, scalabili e sostenibili nei contesti universitari contemporanei.

Quanto al secondo nucleo di ricerca, un risultato particolarmente interessante riguarda il fatto che il feedback del docente continui a rappresentare la fonte di supporto all'apprendimento ritenuta più utile dagli studenti. Questo dato è coerente con quanto emerso nello studio di Grion et al. (2024) e con un'ampia letteratura sul tema, che evidenzia come un feedback docente tempestivo, chiaro e dettagliato possa favorire un apprendimento più profondo e una maggiore

consapevolezza metacognitiva. Attraverso il feedback del docente, infatti, gli studenti hanno la possibilità di individuare con maggiore precisione i propri punti di forza e di debolezza e di sviluppare strategie più efficaci per migliorare le proprie prestazioni (Carless, 2015; Evans, 2013; Nicol, 2010).

Tuttavia, l'elevata utilità attribuita al feedback docente potrebbe essere interpretata anche alla luce di un ulteriore elemento emerso dai dati: le limitate esperienze degli studenti con altre forme di feedback, come quello tra pari o quello automatizzato. È plausibile, infatti, che la maggiore familiarità con il feedback fornito dal docente porti gli studenti a riconoscerlo più facilmente come utile e significativo, mentre le forme di feedback meno sperimentate risultano più difficili da comprendere, definire o valutare in termini di utilità. In altre parole, la preferenza espressa per il feedback docente potrebbe non dipendere esclusivamente da una sua intrinseca superiorità percepita, ma anche da una minore esposizione degli studenti ad altre modalità di feedback, che ne limita la conoscenza e la possibilità di apprezzarne il potenziale contributo ai processi di apprendimento.

Tra le tre modalità di feedback docente analizzate (i.e., scritto, orale, generale in aula), quello scritto è ritenuto dagli studenti il più utile, suggerendo una chiara preferenza per forme di riscontro documentabili, che consentono un'elaborazione approfondita delle informazioni ricevute e un uso strutturato del feedback nel tempo. Tuttavia, sebbene gli studenti riconoscano l'importanza e l'utilità del feedback docente, la loro esperienza effettiva con esso appare frammentaria e discontinua. I dati raccolti mostrano infatti che il 43% degli studenti ha ricevuto feedback dal docente solo qualche volta, mentre il 26.2% dichiara di non averlo mai ricevuto. Questo dato evidenzia una discrepanza significativa tra il valore attribuito al feedback e la sua effettiva implementazione nelle pratiche didattiche universitarie. La letteratura suggerisce che, affinché il feedback sia realmente efficace, deve essere fornito in modo tempestivo e sistematico, così da consentire agli studenti di integrare le informazioni ricevute nel proprio percorso di apprendimento e di applicarle concretamente per migliorare le proprie prestazioni (Winstone et al., 2017). La mancanza di una sistematicità nell'erogazione del feedback docente rappresenta dunque una delle criticità emerse dallo studio, ponendo interrogativi sull'effettiva capacità del sistema universitario di valorizzare il feedback come strumento di regolazione e miglioramento continuo. Questo elemento richiama inoltre la necessità di promuovere percorsi strutturati di formazione dei docenti universitari sulle pratiche valutative e sull'uso pedagogico del feedback. La letteratura evidenzia infatti come il miglioramento delle pratiche di valutazione nell'istruzione superiore richieda lo sviluppo di specifiche competenze valutative da parte dei docenti, spesso ancora considerate una "zona

grigia" della professionalità accademica (Knight, 2002). In questa prospettiva, programmi di *Faculty Development* orientati alla valutazione e al feedback possono rappresentare una leva strategica per favorire pratiche più formative e centrate sull'apprendimento. Come evidenziato da Doria et al. (2025c), percorsi di sviluppo professionale fondati su approcci collaborativi, riflessivi e contestualizzati possono sostenere i docenti nell'integrare in modo più consapevole il feedback e le tecnologie digitali nei processi di insegnamento e valutazione, contribuendo alla costruzione di una cultura della valutazione maggiormente orientata all'apprendimento.

Anche il feedback tra pari viene percepito come utile dagli studenti, sebbene in misura inferiore rispetto a quello docente. Anche nel caso del feedback dei pari, inoltre, la preferenza è per i feedback scritti suggerendo che gli studenti tendano a privilegiare forme di riscontro che possano essere rielaborate e utilizzate in modo strutturato, analogamente a quanto osservato per il feedback del docente. Tuttavia, è probabile che la limitata esperienza degli studenti con il *peer feedback* ne riduca l'efficacia percepita: il 50.7% degli studenti dichiara di averlo ricevuto solo qualche volta, mentre il 31.5% non lo ha mai sperimentato nel proprio percorso accademico. La letteratura evidenzia che il feedback tra pari può rappresentare un'importante risorsa per lo sviluppo di competenze riflessive e autoregolative (Panadero & Lipnevich, 2022; Topping, 2018), ma la sua implementazione nei contesti accademici italiani appare ancora marginale (Doria, 2025). Diversi sono i fattori che possono contribuire alla scarsa diffusione e valorizzazione del *peer feedback*. In primo luogo, la mancanza di formazione specifica dei docenti sulle potenzialità didattiche del peer feedback, e sulle modalità di erogazione e ricezione del feedback tra pari può costituire un ostacolo alla sua piena integrazione nei percorsi di apprendimento (Ajjawi & Boud, 2017). Inoltre, il timore che il giudizio dei pari possa essere meno affidabile rispetto a quello del docente rappresenta un ulteriore elemento di criticità, in quanto gli studenti potrebbero percepire il feedback tra pari come soggettivo o poco attendibile (Nicol, 2021). Per rendere questa pratica più efficace e accettata, sarebbe necessario promuovere strategie didattiche che facilitino l'uso del *peer feedback* in modo strutturato, per esempio attraverso l'impiego di rubriche dettagliate e momenti di revisione tra pari guidati dal docente, così da rafforzare la fiducia degli studenti nella qualità del riscontro ricevuto.

Relativamente al feedback automatizzato, questo viene percepito come utile dagli studenti, sebbene in misura inferiore rispetto a quello fornito dal docente e in misura analoga a quello tra pari. Tuttavia, la percezione degli studenti appare differenziata in funzione della tipologia di strumento analizzato. Nello specifico, i software antiplagio, come Turnitin e Compilatio Studium, risultano gli stru-

menti percepiti come più utili, seguiti dai correttori grammaticali, come Grammarly, mentre i chatbot basati su intelligenza artificiale generativa, come ChatGPT, ricevono le valutazioni più basse. Questi risultati suggeriscono che gli studenti attribuiscono maggiore valore agli strumenti di feedback automatizzato che offrono un riscontro da loro identificabile come chiaro, oggettivo e immediatamente interpretabile su aspetti specifici della propria esperienza di lavoro accademico. Per esempio, è plausibile che i software antiplagio siano apprezzati per la loro capacità di offrire un'analisi dettagliata sull'originalità dei testi, mentre i correttori grammaticali siano considerati utili perché forniscono suggerimenti immediati per migliorare la qualità linguistica e stilistica degli elaborati. Al contrario, la minore valutazione attribuita a strumenti come ChatGPT potrebbe derivare da una diffidenza diffusa o da una limitata conoscenza delle loro potenzialità applicative e delle modalità di interpretazione dei suggerimenti forniti. Analogamente a quanto osservato per il peer feedback, anche il feedback automatizzato risulta essere stato sperimentato solo occasionalmente dagli studenti nella loro esperienza universitaria. I dati raccolti indicano che il 46.5% degli studenti lo ha sperimentato solo qualche volta, mentre il 33.6% non lo ha mai utilizzato. Risultati di analisi in corso sembrerebbero confermare quanto rilevato da Grion e colleghe (2024), secondo cui gli studenti sviluppano percezioni maggiormente positive verso i feedback automatizzati con cui hanno già avuto un'esperienza diretta. Questo fenomeno si inserisce in un quadro più ampio evidenziato dalla letteratura internazionale (Ajjawi & Boud, 2017; Zawacki-Richter et al., 2019), che sottolinea come l'esperienza pregressa con strumenti digitali giochi un ruolo determinante nella valutazione della loro utilità. In altre parole, l'esposizione degli studenti a uno strumento automatizzato porta gli studenti ad attribuire un valore specifico, legato alla propria attività accademica. Limitata familiarità potrebbe spiegare la minore utilità percepita oppure l'aspettativa eccessivamente alta rispetto alle prestazioni degli strumenti di feedback automatizzati.

In tale prospettiva, pare opportuno approfondire questi risultati a partire da successive ricerche empiriche basate non solo sullo studio di indagine, ma anche tramite studi sperimentali con diversi strumenti e livelli di esposizione. Studi longitudinali, volti ad esplorare se e in che modo cambi nel tempo la percezione degli studenti in funzione dell'utilizzo sempre più frequente di strumenti di feedback automatizzato potrebbero far luce sulle relazioni tra familiarità e percezione di utilità del feedback automatizzato. In generale, l'emergente trend di integrazione di tecnologie educative intelligenti nei processi di apprendimento universitario, richiederà un'attenta disamina degli effetti di tali strumenti, in un'ottica di continuo miglioramento dei processi di feedback che bilanci in modo non banale la relazione tra umano e macchina.

Appendice 1
Le prime due sezioni del questionario somministrato a studenti

	Fonte del feedback	Item
S1	Docente	1. I commenti scritti del docente sul lavoro di uno studente mi sono/possono essermi utili per migliorare il mio lavoro.
S2	Docente	2. I commenti orali del docente sul lavoro di uno studente mi sono/possono essermi utili per migliorare il mio lavoro.
S3	Docente	3. Le osservazioni generali fatte dal docente a lezione mi sono/possono essermi utili per migliorare il mio lavoro.
S4	Pari	4. I commenti scritti dei compagni in merito ad un lavoro svolto da uno studente durante le lezioni in corso, mi sono/possono essermi utili per migliorare il mio lavoro.
S5	Pari	5. I commenti orali dei compagni in merito ad un lavoro svolto da uno studente durante le lezioni in corso, mi sono/possono essermi utili per migliorare il mio lavoro.
S6	Pari	6. Le osservazioni generali fatte dai compagni, durante le lezioni, mi sono/possono essermi utili per migliorare il mio lavoro.
S7	Pari	7. I commenti fatti dai compagni in situazioni diverse da quelle della lezione mi sono/possono essermi utili per migliorare il mio lavoro.
S8	Computer	8. Le analitiche fornite dai sistemi come LMS (Moodle, Blackboard, Olat, Teams, ecc.) (come i miei log, completamento di attività o le barre di progresso) mi sono/possono essermi utili per migliorare il mio lavoro.
S9	Computer	9. I feedback automatici virtuali (come i punteggi dei quiz e il completamento dei compiti) mi sono/possono essermi utili per migliorare il mio lavoro.
S10	Computer	10. I feedback da bacheche e grafiche su altre applicazioni (come Annoto, Perusall, Mentimeter, Wooclap) integrate nella piattaforma LMS (Moodle, Blackboard, Olat, Teams, ecc.), mi sono/possono essermi utili per migliorare il mio lavoro.
S11-T	Computer	11. Guarda il seguente strumento: ChatGPT. Si tratta di una chatbot in grado di scrivere contenuti utilizzabili in numerosi contesti, effettuare una traduzione automatica, fornire informazioni e generare risposte in tempo reale. Quanto pensi che potrebbe aiutarti per il tuo lavoro?
S12-T	Computer	12. Guarda il seguente strumento: Grammarly. Si tratta di una piattaforma per il controllo ortografico e la correttezza grammaticale del testo. Quanto pensi che potrebbe aiutarti per il tuo lavoro?
S13-T	Computer	13. Guarda i seguenti strumenti: Compilatio Studium o Turnitin. Si tratta di strumenti in grado di rilevare similitudini con altri testi/contenuti al fine di evitare il plagio. Quanto pensi che potrebbero aiutarti per il tuo lavoro?

S14	Computer	14. I feedback ottenuti da sistemi di analisi del testo (che mi danno indicazioni sulla qualità dei miei testi, per esempio analizzando chiarezza, uso di fonti, originalità nella trattazione di un tema, ecc.) in modalità di grafiche oppure di raccomandazioni, mi sono/possono essermi utili per migliorare il mio lavoro.
S15	Computer	15. I feedback ottenuti da sistemi di analisi degli interventi sul forum – che mi mostrano grafiche sulla partecipazione e la collaborazione, oppure mi danno indicazioni sul miglioramento della mia partecipazione/intervento – mi sono/possono essermi utili per migliorare il mio lavoro.

Sezione 1 – Situazioni e strumenti di feedback e relativi item del questionario
Note. Situazioni (S) e Strumenti (T) di feedback

	Esperienza di feedback	Item
E1	Feedback docente	1. Nel corso della mia carriera universitaria, ho frequentato corsi dove il docente forniva feedback.
E2	Feedback tra pari	2. Nel corso della mia carriera universitaria, ho frequentato corsi dove il docente stimolava esplicitamente processi di feedback tra pari.
E3	Auto-feedback	3. Nel corso della mia carriera universitaria, ho frequentato corsi dove il docente stimolava esplicitamente processi di auto-feedback.
E4	Feedback automatizzato	4. Nel corso della mia carriera universitaria, ho frequentato corsi dove si sono utilizzati sistemi digitali con forme di feedback automatici (es: quiz online con risposte di orientamento, barra di avanzamento del corso, grafici interattivi sulle attività completate o sulle tue competenze).

Sezione 2 – Esperienza di feedback e relativi item del questionario

Riferimenti bibliografici

Ajjawi, R., & Boud, D. (2017). Researching feedback dialogue: An interactional analysis approach. *Assessment & Evaluation in Higher Education, 42*(2), 252-265.

Bañeres, D., Rodríguez, M. E., Guerrero-Roldán, A. E., & Karadeniz, A. (2020). An early warning system to detect at-risk students in online higher education. *Applied Sciences, 10*(13), 4427.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31.

Bonaiuti, G., & Dipace, A. (2021). *Insegnare e apprendere in aula e in rete: Per una didattica blended efficace.* Carocci.

Bozzi, M., Raffaghelli, J. E., & Zani, M. (2021). Peer learning as a key component of an integrated teaching method: Overcoming the complexities of physics teaching in large size classes. *Education Sciences*, *11*(2), 67.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101.

Brusilovsky, P., & Peylo, C. (2003). Adaptive and intelligent Web-based educational systems. *International Journal of Artificial Intelligence in Education, 13*(2-4), 159-171.

Carless, D. (2015). Exploring learning-oriented assessment processes. *Higher Education*, *69*, 963-976.

Carless, D. (2019). Feedback loops and the longer-term: towards feedback spirals. *Assessment & Evaluation in Higher Education, 44*(5), 705-714.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315-1325.

Crow, T., Luxton-Reilly, A., & Wuensche, B. (2018). Intelligent tutoring systems for programming education: a systematic review. In *Proceedings of the 20th Australasian Computing Education Conference* (pp. 53-62).

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences, 116*(39), 19251-19257.

Ding, H. (2023). Qualitative Comparative Analysis: Search Target, Reflection on the Top-Down Approach, and Introduction of the Bottom-Up Approach. *International Journal of Qualitative Methods*, 22.

Doria, B. (2025). *Ripensare la valutazione degli apprendimenti all'università. Dalla formazione dei docenti alle pratiche valutative d'aula*. Lecce: Pensa MultiMedia.

Doria, B., & Grion, V. (2020). L'autovalutazione nel contesto universitario: Una revisione sistematica della letteratura. *Form@re - Open Journal per la formazione in rete, 20*(1), 78-92.

Doria, B., & Picasso, F. (2024). Alternative assessment and technology-enhanced assessment practices: Research to inform faculty development processes. *QWERTY-Interdisciplinary Journal of Technology, Culture and Education, 19*(1), 52-71.

Doria, B., Foschi, L. C., Raffaghelli, J. E., & Grion, V. (2025a). Percezioni ed esperienze degli studenti universitari rispetto al feedback docente, tra pari e automatico. *Italian Journal of Educational Research, (34)*, 135-149.

Doria, B., Foschi, L. C., Slaviero, G., Zaggia, C., & Grion, V. (2025b). Feedback e Feedback Automatizzato: Pratiche e percezioni dei docenti universitari. *Research Trends In Humanities, 12*, 25-42.

Doria, B., Grion, V., & Paccagnella, O. (2023). Pratiche valutative nelle università italiane: Una ricerca esplorativa a livello nazionale. *Italian Journal of Educational Research, 30*, 129-143.

Doria, B., Grion, V., Picasso, F., & Li, L. (2025c). Faculty assessment development in higher education: the CRAFT framework emerging from a systematic literature review. *Assessment & Evaluation in Higher Education*, 1-25. https://doi.org/10.1080/-02602938.2025.2584121

Ergas, O. (2017). The Curriculum of Embodied Perception. In *Reconstructing 'Educa-*

tion' through Mindful Attention. Palgrave Macmillan, London. https://doi.org/-10.1057/978-1-137-58782-4_7

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research, 83*(1), 70-120.

Facer, K., & Selwyn, N. (2021). Digital technology and the futures of education: Towards 'Non-Stupid' optimism. *Futures of Education initiative, UNESCO*.

Foschi, L. C., Doria, B., Laici, C., Gratani, F., Screpanti, L., Fiorentino, M. G., Rossi, P. G., Giannandrea, L., Grion, V., & Montone, A. (2024). AI e Feedback. Interazione tra agenti umani e artificiali per valutare prove scritte in ambito universitario. *Education Sciences & Society, 2*, 142-156.

Grion, V., & Cesareni, D. (2016). Multiplicity, fluidity, dialogue and sharing: Keywords to understand the complex dynamics between human learning and technology. *QWERTY – Interdisciplinary Journal of Technology, Culture and Education, 11*(1), 5-10.

Grion, V., & Serbati, A. (2019). *Valutazione sostenibile e feedback nei contesti universitari. Prospettive emergenti, ricerche e pratiche*. Pensa MultiMedia.

Grion, V., Raffaghelli, J., Doria, B., & Serbati, A. (2024). Students' perceptions on different sources of self-feedback. *Educational Research and Evaluation*, 1-23.

Grion, V., Serbati, A., Doria, B., & Nicol, D. (2021). Rethinking assessment and feedback practices in higher education: A review of recent literature. *Innovations in Education and Teaching International, 58*(4), 405-416.

Guskey, T. R. (2019). *Get set, go! Implementing successful reforms in grading and reporting*. Solution Tree.

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 6300-6308).

Knight, P.T. (2002). The Achilles' Heel of Quality: The assessment of student learning. *Quality in Higher Education,* 8(1), 107-115. DOI: 10.1080/13538320220127506

Lancho, M. S., Hernández, M., Paniagua, Á. S. E., Encabo, J. M. L., & de Jorge-Botana, G. (2018). *Using semantic technologies for formative assessment and scoring in large courses and MOOCs*.

Lipnevich, A. A., Panadero, E., Gjicali, K., & Fraile, J. (2021). What's on the syllabus? An analysis of assessment criteria in first-year courses across US and Spanish universities. *Educational Assessment, Evaluation and Accountability, 33*(3), 675–699.

Molenaar, I. (2022). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. *Computers and Education: Artificial Intelligence*, *3*, 100070.

Nguyen, J., Sánchez-Hernández, G., Armisen, A., Agell, N., Rovira, X., & Angulo, C. (2018). A linguistic multi-criteria decision-aiding system to support university career services. *Applied Soft Computing Journal, 67*, 933-940.

Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education, 35*(5), 501-517.

Nicol, D. (2021). Guiding learning by activating students' inner feedback. *Times Higher Education*.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

Orsmond, P., Merry, S., & Reiling, K. (2005). Biology students' utilization of tutors' formative feedback: A qualitative interview study. *Assessment & Evaluation in Higher Education, 30*(4), 369-386.

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, *35*, 100416.

Pauli, M., & Ferrell, G. (2020). *The future of assessment: five principles, five targets for 2025*.

Picasso, F., Doria, B., Grion, V., Venuti, P., & Serbati, A. (2023). What technology-enhanced assessment and feedback practices do Italian academics declare in their syllabi? Analysis and reflections to support academic development. In G. Fulantelli, D. Burgos, G. Casalino, M. Cimitile, G. Lo Bosco, & D. Taibi (Eds.), *Higher education learning methodologies and technologies online* (pp. 267-279). Springer.

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education, 35*(3), 277-289.

Raffaghelli, J. E. (2024). *Post-digital scholarship: Professionalità accademica e trasformazione digitale in università*. Pensa MultiMedia.

Raffaghelli, J. E., Ferrarelli, M., & Rodríguez, N. L. (2025). Slowness as Postdigital Positionality in the era of generative AI: A Conversation. *Postdigital Science and Education*, 7(4), 1224-1249.

Raffaghelli, J., Ghislandi, P., Sancassani, S., Canal, L., Micciolo, R., Balossi, B., & Zani, M. (2018). Integrating MOOCs in physics preliminary undergraduate education: Beyond large size lectures. *Educational Media International, 55*(4), 301-316.

Ranieri, M., Raffaghelli, J., & Pezzati, F. (2018). Digital resources for faculty development in e-learning: A self-paced approach for professional learning. *Italian Journal of Educational Technology, 26*(1), 104-118.

Redecker, C., & Punie, Y. (2017). *European framework for the digital competence of educators: DigCompEdu.* Publications Office of the European Union.

Rodríguez, M. E., Guerrero-Roldán, A. E., Bañeres, D., & Karadeniz, A. (2022). An intelligent nudging system to guide online learners. *The International Review of Research in Open and Distributed Learning, 23*(1), 41-62.

Rodríguez, M. E., Raffaghelli, J. E., Bañeres, D., Guerrero-Roldán, A. E., & Crudele, F. (2024). Exploring Higher Education Students' Experience with AI-powered Educational Tools: The Case of an Early Warning System . *Formazione & Insegnamento*, *22*(1), 74-84. https://doi.org/10.7346/-fei-XXII-01-24_09

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144.

Serbati, A., Grion, V., & Fanti, M. (2019). Caratteristiche del peer feedback e giudizio valutativo in un corso universitario blended. *Giornale Italiano della Ricerca Educativa, 12*(numero speciale), 115-137.

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, 33(1), 1-19.

Sim, G.R., Holifield, P., & Brown, M. (2004). Implementation of computer assisted assessment: lessons from the literature. *Research in Learning Technology, 12*, 215-229.

Sullivan, J. V. (2018). Learning and Embodied Cognition: A Review and Proposal. *Psychology Learning & Teaching, 17*(2), 128-143. https://doi.org/10.1177/1475725-717752550

Tonelli, D., Grion, V., & Serbati, A. (2018). L'efficace interazione fra valutazione e tecnologie: evidenze da una rassegna sistematica della letteratura. *Italian Journal of Educational Technology, 26*(3), 6-23. doi: 10.17471/2499-4324/1028

Topping, K. (2018). *Using peer assessment to inspire reflection and learning*. Routledge.

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*, 17-37.

Yoo, Y., Lee, H., Jo, I. H., & Park, Y. (2015). Educational dashboards for smart learning: Review of case studies. In G. Chen, V. Kumar, Kinshuk, R. Huang & S. Kong, (Eds.) *Emerging issues in smart learning* (pp. 145-155). Springer Berlin Heidelberg.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–Where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1-27.

I.3
L'intelligenza artificiale per il feedback formativo in matematica: specificità epistemologiche, modelli e progettazione nei problemi aperti per la formazione universitaria

Antonella Montone
Professoressa associata, Università degli Studi di Bari Aldo Moro, antonella.montone@uniba.it
Michele Giuliano Fiorentino
Assegnista di ricerca, Università degli Studi di Bari Aldo Moro, michele.fiorentino@uniba.it

Abstract

Il contributo analizza il ruolo del feedback automatizzato nei contesti universitari che coinvolgono la Matematica, evidenziandone le specificità epistemologiche e didattiche. Nell'ambito del progetto PRIN "AI&F", l'attenzione è posta non tanto sulle potenzialità generali dell'intelligenza artificiale, quanto sulle condizioni che ne rendono possibile un utilizzo efficace in presenza di compiti aperti e strategie risolutive plurali. A partire da una rilettura dei modelli di feedback in relazione allo specifico ambito della didattica della matematica, il lavoro discute criticità e vincoli legati alla natura del discorso matematico, alla molteplicità dei registri semiotici e alla varietà delle argomentazioni. Viene inoltre proposto un modello interpretativo del feedback mediato dall'IA che integra dimensioni semiotiche e argomentative. I risultati evidenziano come l'efficacia del feedback automatizzato dipenda dalla progettazione didattica e dalla capacità di integrare l'interazione tra docente e sistema all'interno di processi di costruzione del significato.

Parole chiave: Didattica della Matematica; intelligenza artificiale generativa; feedback; valutazione; università.

This paper examines the role of automated feedback in university contexts involving Mathematics, highlighting its epistemological and didactic specificities. Within the framework of the PRIN "AI&F" project, the focus is not so much on the general potential of artificial intelligence, but rather on the conditions that enable its effective use in the presence of open-ended tasks and multiple solution strategies. Drawing on a reinterpretation of feedback models through the lens of mathematics education, the study discusses critical issues and constraints related to the nature of mathematical discourse, the multiplicity of semiotic registers, and the variety of argumentation forms. It also proposes an interpretative

model of AI-mediated feedback that integrates semiotic and argumentative dimensions. The findings show that the effectiveness of automated feedback depends on instructional design and on the ability to integrate the interaction between teacher and system within processes of meaning construction.

Keywords: Mathematics education; generative AI; feedback; assessment; higher education

1. Introduzione

Negli ultimi anni, il tema del feedback ha assunto un ruolo centrale nei processi di insegnamento e apprendimento nell'istruzione superiore, in particolare in relazione alla crescente numerosità degli studenti e alla conseguente difficoltà nel garantire forme di feedback tempestive, personalizzate e pedagogicamente efficaci (Winstone & Carless, 2019; Lipnevich & Panadero, 2021). In tale scenario, l'intelligenza artificiale (IA) è stata progressivamente considerata come una possibile risorsa per rendere più sostenibili le pratiche valutative, soprattutto nei contesti caratterizzati da compiti complessi e da un elevato carico di correzione (Deeva et al., 2021; Corbin et al., 2025). Tuttavia, quando si considera il dominio della matematica, emergono criticità specifiche che rendono non immediatamente trasferibili i modelli generali di feedback automatizzato. La natura del sapere matematico, caratterizzata da pluralità di strategie, uso di linguaggi simbolici e necessità di giustificazione, introduce vincoli epistemologici che incidono profondamente sulla progettazione del feedback. L'IA quindi si configura come una risorsa potenzialmente decisiva per rendere sostenibili pratiche valutative di tipo formativo, soprattutto in presenza di compiti aperti e cognitivamente complessi.

Nel quadro teorico del progetto PRIN "Artificial Intelligence & Feedback for Effective Learning (AI&F)", il feedback viene concepito non come semplice trasmissione di informazioni correttive, ma come un processo dialogico, generativo e trasformativo, in grado di attivare cicli iterativi di riflessione, revisione e costruzione del significato (Sadler, 2010; Nicol, 2010; Carless & Boud, 2018). In questa prospettiva, l'IA non sostituisce il docente, ma si configura come partner cognitivo, inserito in un sistema di interazione tra agente umano e agente artificiale (Bearman & Ajjawi, 2023). Se tali considerazioni sono valide in generale, assumono una rilevanza ancora maggiore nel dominio della matematica, disciplina caratterizzata da specifiche esigenze epistemologiche e semiotiche. In particolare, discipline come la matematica presentano sfide complesse, legate alla varietà delle strategie risolutive, al linguaggio specialistico e alle diverse rap-

presentazioni (algebriche, grafiche, geometriche). Per questo, i modelli di machine learning devono essere in grado di riconoscere approcci differenti, anche non convenzionali, comprendere la sintassi matematica e valutare la correttezza e la coerenza del ragionamento.

Il sistema deve inoltre adattarsi alle diverse modalità espressive degli studenti e favorire un ambiente digitale che supporti la scrittura e l'interpretazione delle soluzioni.

Il presente contributo, in continuità con lo studio condotto presso la sede di Macerata, si propone di analizzare le specificità del feedback automatizzato in matematica, sperimentando gli strumenti elaborati e adattandoli alla specificità disciplinare con particolare attenzione ai problemi aperti in contesti universitari ad alta numerosità, integrando prospettive provenienti dalla didattica della matematica e dalle tecnologie dell'IA. L'obiettivo è garantire una valutazione accurata, equa e formativa, capace di fornire feedback mirati e personalizzati.

In questo quadro, l'intelligenza artificiale non si configura soltanto come uno strumento a supporto della valutazione, ma come un elemento che interviene nella struttura stessa dell'interazione didattica, contribuendo a ridefinire le modalità attraverso cui il feedback viene prodotto, interpretato e utilizzato dagli studenti.

2. Quadro teorico

La matematica si distingue per una natura epistemologica in cui il sapere è costruito attraverso relazioni, inferenze e sistemi di giustificazione. Di conseguenza, il feedback non può essere ridotto alla verifica della correttezza del risultato, ma deve intervenire sulla qualità del processo di costruzione del significato (Pehkonen, 1997; Sfard, 2008).

La progettazione di sistemi di feedback automatizzato in matematica richiede di operare su almeno tre piani distinti: quello sintattico, relativo alla manipolazione simbolica; quello semantico, relativo al significato matematico; e quello argomentativo, relativo alla costruzione del ragionamento. La difficoltà non risiede soltanto nella loro individuazione, ma nella capacità del sistema di coordinarli all'interno di un'unica interpretazione della risposta dello studente. Questa prospettiva si colloca in continuità con la teoria del discorso matematico (Sfard, 2008), secondo cui apprendere matematica significa partecipare a pratiche discorsive specifiche, caratterizzate da regole di validazione condivise.

Un ulteriore contributo fondamentale è offerto dalla teoria dei registri di rappresentazione semiotica (Duval, 2006). La comprensione matematica implica

la capacità di coordinare registri diversi (simbolico, grafico, linguistico), e molte difficoltà derivano proprio dall'incapacità di operare conversioni tra tali registri. Il feedback efficace deve quindi agire come mediatore semiotico, supportando il passaggio tra rappresentazioni e rendendo esplicite le relazioni sottostanti.

In questa prospettiva, il feedback assume una funzione di mediazione che non si limita alla restituzione di informazioni, ma interviene nella relazione tra soggetto, contenuto e rappresentazioni, contribuendo alla costruzione del significato matematico. Tale funzione diventa particolarmente rilevante nei contesti in cui il feedback è mediato da sistemi di intelligenza artificiale, che introducono nuove modalità di interazione e nuovi vincoli interpretativi.

I problemi aperti costituiscono un contesto privilegiato per lo sviluppo del pensiero matematico, ma introducono una complessità valutativa significativa. In primo luogo, essi definiscono uno spazio di soluzioni aperto, in cui non esiste una corrispondenza univoca tra problema e risposta (Pehkonen, 1997). In accordo con la teoria antropologica di Chevallard (1999), ciò implica la presenza di una pluralità di prasseologie, ossia combinazioni di tecniche, giustificazioni e teorie, mentre la teoria delle situazioni didattiche (Brousseau, 1997) evidenzia il ruolo del milieu nella costruzione del significato.

In questo quadro, il feedback deve essere progettato in modo da non interrompere il processo esplorativo, evitando di fornire soluzioni premature. Un ulteriore elemento di complessità riguarda l'argomentazione. Come evidenziato da Balacheff (1999), le giustificazioni degli studenti si collocano lungo un continuum che va da forme empiriche a forme deduttive. Il feedback deve sostenere il passaggio verso livelli più avanzati di rigore, senza svalutare le forme iniziali di ragionamento. Infine, dal punto di vista cognitivo (Dubinsky & McDonald, 2001), il feedback deve essere allineato alla fase di sviluppo concettuale dello studente, supportando il passaggio da azioni a processi e oggetti matematici.

Alla luce delle specificità disciplinari, i modelli di feedback automatizzato possono essere reinterpretati in chiave didattica. Il feedback valutativo, centrato sulla correttezza, appare insufficiente in contesti aperti, in quanto non supporta la costruzione del significato (Black & Wiliam, 2009). Il feedback diagnostico consente di individuare errori specifici, ma richiede una modellizzazione esplicita delle strategie e delle difficoltà. Il feedback elaborativo fornisce spiegazioni e connessioni, ma può risultare eccessivamente direttivo. Il feedback generativo e dialogico rappresenta il livello più avanzato: esso agisce sul discorso matematico dello studente (Sfard, 2008), promuovendo riflessione, argomentazione e revisione. In termini vygotskiani, può essere interpretato come forma di scaffolding nella zona di sviluppo prossimale, a condizione che non diventi eccessivamente prescrittivo.

3. Tecniche di intelligenza artificiale per il feedback matematico

Le tecnologie di intelligenza artificiale applicate al feedback matematico si collocano lungo un continuum che riflette differenti modalità di trattamento dell'informazione e diversi livelli di aderenza alla natura del sapere matematico. Dal punto di vista didattico, tali approcci non sono equivalenti, poiché introducono vincoli specifici nella progettazione del feedback e nella sua efficacia rispetto ai processi di apprendimento.

Gli approcci simbolici, tipicamente implementati attraverso sistemi di algebra computazionale, garantiscono un elevato livello di rigore formale e risultano particolarmente efficaci nella verifica della correttezza sintattica e procedurale delle soluzioni. Tuttavia, essi mostrano limiti evidenti nel cogliere la dimensione semantica e discorsiva del ragionamento matematico, riducendo spesso il feedback a una valutazione binaria della correttezza, senza restituire indicazioni significative sui processi sottostanti.

Diversamente, gli approcci basati su embedding e rappresentazioni vettoriali consentono di analizzare la similarità tra testi e di individuare relazioni semantiche tra risposte degli studenti e modelli di riferimento. Sebbene tali strumenti rappresentino un avanzamento rispetto ai sistemi puramente simbolici, essi risultano fortemente sensibili alla formulazione linguistica e possono incontrare difficoltà nel riconoscere equivalenze matematiche espresse attraverso registri diversi o strategie non convenzionali.

I modelli generativi, e in particolare i Large Language Models, introducono una dimensione ulteriore, permettendo la produzione di feedback articolati, contestuali e dialogici. Essi sono in grado di simulare forme di discorso matematico (Sfard, 2008) e di sostenere l'interazione con lo studente attraverso domande, suggerimenti e riformulazioni. Tuttavia, il loro funzionamento basato su correlazioni statistiche comporta una criticità rilevante: la mancanza di garanzia sulla correttezza matematica del contenuto prodotto, che può risultare plausibile dal punto di vista linguistico ma non fondato sul piano epistemico.

Alla luce di queste considerazioni, emerge la necessità di orientarsi verso soluzioni ibride, in cui diverse tecnologie vengano integrate in modo complementare. In particolare, appare promettente l'integrazione tra validazione simbolica, analisi semantica e generazione linguistica, al fine di bilanciare rigore formale e ricchezza comunicativa. In tali sistemi, la correttezza matematica può essere garantita da moduli simbolici, mentre la dimensione interpretativa e dialogica può essere affidata a modelli generativi, permettendo così la costruzione di feedback più aderenti alle esigenze della didattica della matematica.

Questa articolazione mette in evidenza come le tecnologie di intelligenza ar-

tificiale non siano neutre rispetto alla didattica, ma incorporino specifiche modalità di interpretazione e restituzione delle produzioni degli studenti. Di conseguenza, la scelta degli strumenti influisce non solo sulla qualità del feedback, ma anche sulle forme di interazione che si sviluppano tra studente, contenuto matematico e sistema.

4. Criticità del feedback automatizzato in matematica

L'introduzione dell'intelligenza artificiale nei processi di feedback in matematica solleva una serie di criticità che non possono essere interpretate esclusivamente come limiti tecnici dei sistemi, ma devono essere comprese alla luce della natura epistemologica della disciplina. In particolare, la matematica pone vincoli specifici alla progettazione del feedback, legati alla complessità del discorso matematico, alla pluralità delle strategie risolutive e alla necessità di giustificazione rigorosa.

Una prima criticità riguarda la difficoltà nel riconoscere equivalenze matematiche profonde. In matematica, infatti, soluzioni formalmente differenti possono essere concettualmente equivalenti, mentre soluzioni apparentemente simili possono nascondere errori significativi. I sistemi automatici, soprattutto quelli basati su similarità superficiali o su modelli statistici del linguaggio, faticano a cogliere tali distinzioni, con il rischio di validare risposte scorrette o penalizzare strategie corrette ma non convenzionali. Questo problema è strettamente connesso alla natura del sapere matematico come sistema di relazioni e giustificazioni (Sfard, 2008).

Una seconda criticità è legata all'ambiguità semiotica, ovvero alla presenza di molteplici registri di rappresentazione (Duval, 2006). Gli studenti possono esprimere uno stesso concetto matematico attraverso linguaggi diversi (simbolico, grafico, verbale), e la comprensione richiede la capacità di coordinare tali registri. I sistemi di intelligenza artificiale, tuttavia, tendono a operare su rappresentazioni parziali e possono incontrare difficoltà nell'integrare informazioni provenienti da registri diversi, riducendo la qualità del feedback.

Un ulteriore elemento problematico riguarda i modelli generativi, i quali, pur essendo in grado di produrre feedback articolati e contestuali, non garantiscono la correttezza matematica delle risposte. Come evidenziato in letteratura recente, tali modelli possono generare contenuti plausibili dal punto di vista linguistico ma non fondati sul piano epistemico (Corbin et al., 2025). Questo introduce un rischio significativo nei contesti educativi, in quanto lo studente può essere indotto ad accettare indicazioni non corrette.

Quest'ultimo aspetto evidenzia una tensione strutturale tra la natura linguistica dei modelli generativi e la specificità del discorso matematico. In particolare, mentre i sistemi di IA operano prevalentemente su correlazioni testuali, la matematica richiede relazioni di tipo logico e giustificativo, che non sempre risultano pienamente esplicitate a livello linguistico. Ne deriva che un feedback apparentemente coerente sul piano discorsivo può risultare debole o problematico sul piano matematico.

Dal punto di vista didattico, emerge inoltre il rischio di over-scaffolding. Un feedback eccessivamente guidato può ridurre l'impegno cognitivo dello studente, limitando la sua capacità di esplorazione autonoma e di costruzione del significato. In contesti di problemi aperti, questo rischio è particolarmente rilevante, poiché il valore formativo dell'attività risiede proprio nella possibilità di sviluppare strategie personali e argomentazioni autonome (Schoenfeld, 1985).

Infine, l'introduzione dell'IA incide sui processi di trasposizione didattica (Chevallard, 1999) e sul contratto didattico (Brousseau, 1997). La presenza di un agente artificiale modifica implicitamente le aspettative degli studenti rispetto alla natura del sapere matematico e al ruolo del docente, rendendo necessaria una progettazione consapevole delle modalità di integrazione dell'IA nei contesti educativi.

5. Un modello teorico: The AI-Mediated Mathematical Feedback Cycle (AIM-FC)

Alla luce delle criticità e delle potenzialità emerse, si propone il modello AI-Mediated Mathematical Feedback Cycle (AIM-FC), che interpreta il feedback non come un evento puntuale, ma come un processo iterativo e trasformativo.

Il ciclo prende avvio dalla produzione dello studente, intesa come manifestazione di un processo di costruzione del significato. Tale produzione viene successivamente interpretata dal sistema di intelligenza artificiale, che ne analizza le caratteristiche sintattiche, semantiche e discorsive. Questa fase di interpretazione rappresenta un passaggio cruciale, in quanto determina la qualità del feedback che verrà generato.

Il feedback prodotto dal sistema si configura come uno stimolo alla riflessione e alla rielaborazione, piuttosto che come una semplice correzione. In questa prospettiva, esso attiva una fase di mediazione semiotica (Duval, 2006), in cui lo studente è chiamato a riconsiderare la propria produzione, mettendo in relazione diversi registri di rappresentazione.

Parallelamente, il ciclo sostiene lo sviluppo argomentativo, favorendo il passaggio da forme intuitive o empiriche di giustificazione a forme più strutturate

e formalizzate (Balacheff, 1999). Questo processo è coerente con la visione del discorso matematico come pratica sociale e comunicativa (Sfard, 2008).

Un ulteriore elemento del ciclo è rappresentato dal raffinamento concettuale, attraverso il quale lo studente riorganizza le proprie conoscenze, integrando nuove informazioni e rivedendo le proprie strategie. Il ciclo si conclude con una revisione della produzione iniziale, che rappresenta un avanzamento nel processo di costruzione del significato.

Il modello evidenzia due assi principali di trasformazione. Il primo è l'asse semiotico, relativo alla coordinazione dei registri di rappresentazione (Duval, 2006); il secondo è l'asse argomentativo, relativo allo sviluppo della giustificazione (Balacheff, 1999). In questa prospettiva, il feedback mediato dall'IA si configura come un dispositivo capace di attivare processi trasformativi profondi, a condizione che sia progettato in modo coerente con la natura del sapere matematico.

In questa prospettiva, il feedback non è interpretato come un prodotto generato dal sistema, ma come un processo emergente dall'interazione tra studente, intelligenza artificiale e contesto didattico. Il modello descrive dunque un dispositivo dinamico, nel quale il significato si costruisce attraverso cicli successivi di produzione, interpretazione e rielaborazione. Tale carattere ciclico consente di evitare una lettura lineare del feedback e di valorizzare, invece, la sua funzione regolativa nei processi di apprendimento matematico.

6. Disegno metodologico e contesto di sperimentazione

Nel quadro del progetto PRIN "AI&F", l'unità di ricerca dell'Università di Bari (WP3) ha sviluppato una fase di sperimentazione focalizzata sull'adattamento dei modelli di feedback automatizzato al dominio della matematica.

La sperimentazione è stata condotta in contesti universitari caratterizzati da elevata numerosità, nei quali il problema della sostenibilità del feedback risulta particolarmente rilevante (Winstone & Carless, 2019; Lipnevich & Panadero, 2021). I compiti proposti sono stati progettati privilegiando problemi aperti, in grado di rendere visibili strategie e forme di argomentazione (Pehkonen, 1997; Schoenfeld, 1985).

Il disegno metodologico ha previsto un processo iterativo articolato in tre fasi: progettazione dei compiti e dell'interazione con l'IA, raccolta delle produzioni degli studenti, analisi qualitativa delle interazioni. Quest'ultima è stata condotta alla luce della teoria del discorso matematico (Sfard, 2008), dei registri semiotici (Duval, 2006) e della valutazione formativa (Black & Wiliam, 2009).

In questo senso, l'analisi non è stata orientata alla valutazione delle prestazioni del sistema, ma alla comprensione dei processi attivati dall'interazione con esso. L'attenzione è stata posta in particolare sul modo in cui il feedback generato dall'IA interviene nella riorganizzazione delle produzioni degli studenti e nella progressiva esplicitazione dei significati matematici.

L'IA è stata interpretata come mediatore del processo di feedback (Carless & Boud, 2018; Bearman & Ajjawi, 2023), con particolare attenzione alla progettazione dell'interazione, al fine di favorire processi di rielaborazione e sviluppo argomentativo.

7. Fasi della sperimentazione e protocollo metodologico

La sperimentazione è stata progettata con l'obiettivo di analizzare il ruolo dell'intelligenza artificiale generativa nei processi di feedback in matematica, con particolare riferimento a compiti aperti in contesti universitari ad alta numerosità. In coerenza con un approccio iterativo di tipo design-based, il protocollo ha previsto una strutturazione in tre fasi successive, finalizzate a osservare l'evoluzione delle produzioni degli studenti in relazione all'interazione con il sistema.

Nella prima fase, agli studenti è stato proposto un problema matematico aperto, progettato per rendere visibili strategie risolutive, scelte rappresentative e forme di argomentazione (Pehkonen, 1997; Schoenfeld, 1985). Le risposte prodotte in questa fase sono state considerate come base-line, ovvero come espressione iniziale dei processi di costruzione del significato.

La seconda fase ha previsto l'interazione con un sistema di intelligenza artificiale generativa, guidata attraverso prompt progettati dal docente. Tali prompt non erano orientati alla fornitura della soluzione, ma alla generazione di feedback formativi, finalizzati a sollecitare la riflessione, l'esplicitazione dei passaggi impliciti e la revisione del ragionamento. In questa fase, l'IA è stata configurata come mediatore del processo di feedback (Carless & Boud, 2018; Bearman & Ajjawi, 2023), in grado di attivare una prima rielaborazione individuale.

Nella terza fase, gli studenti sono stati invitati a produrre una versione revisionata della propria soluzione, documentando, ove richiesto, l'interazione con il sistema. Le produzioni iniziali, i feedback generati e le risposte rielaborate sono stati successivamente analizzati qualitativamente e utilizzati come base per una discussione collettiva guidata dal docente. Tale momento ha svolto una funzione di istituzionalizzazione dei significati matematici, permettendo di ricondurre le elaborazioni individuali a un quadro condiviso (Sfard, 2008).

L'analisi dei dati è stata condotta secondo una prospettiva qualitativa, focalizzata su tre dimensioni principali. La prima riguarda la trasformazione delle strategie risolutive, osservata attraverso il confronto tra la produzione iniziale e la versione rielaborata dopo l'interazione con l'IA. La seconda concerne l'evoluzione dei registri semiotici utilizzati dagli studenti, con particolare attenzione alla capacità di coordinare rappresentazioni verbali, simboliche e grafiche (Duval, 2006). La terza dimensione riguarda lo sviluppo delle forme di argomentazione, inteso come passaggio da giustificazioni intuitive o empiriche a forme più esplicite e strutturate di ragionamento matematico (Balacheff, 1999).

Particolare attenzione è stata posta all'interazione tra feedback generato dall'IA e successiva rielaborazione dello studente, interpretata come indicatore dell'attivazione di un feedback loop (Carless & Boud, 2018).

8. Implicazioni didattiche

L'integrazione dell'intelligenza artificiale nei processi di feedback in matematica implica una ridefinizione significativa dei ruoli all'interno del sistema didattico. In primo luogo, il docente assume il ruolo di progettista del feedback, responsabile non solo della definizione dei compiti, ma anche della configurazione delle modalità di interazione con il sistema di IA. Ciò richiede competenze specifiche nella progettazione didattica e nella comprensione dei limiti e delle potenzialità delle tecnologie utilizzate.

Parallelamente, lo studente è chiamato a svolgere un ruolo attivo nel processo di apprendimento, sviluppando competenze di feedback literacy, ovvero la capacità di interpretare, valutare e utilizzare il feedback per migliorare le proprie prestazioni (Carless & Boud, 2018). Questo implica un cambiamento nella concezione del feedback, da strumento di correzione a risorsa per la costruzione del significato.

L'uso dell'IA richiede inoltre la progettazione di ambienti di apprendimento in cui l'interazione tra agente umano e agente artificiale sia esplicitamente regolata. In tali ambienti, il feedback automatizzato deve essere integrato con momenti di discussione e riflessione guidata dal docente, al fine di garantire la qualità del processo di apprendimento.

Un ulteriore aspetto riguarda la progettazione delle rubriche valutative e dei criteri di analisi delle risposte. In contesti matematici, tali strumenti devono essere in grado di valorizzare non solo la correttezza del risultato, ma anche la qualità del ragionamento e la capacità di argomentazione (Black & Wiliam, 2009).

In questo senso, l'IA può supportare il docente nella gestione del feedback, ma non può sostituire il giudizio professionale.

Infine, appare necessario promuovere una riflessione critica sull'uso delle tecnologie, evitando approcci deterministici e riconoscendo che l'efficacia dell'IA dipende dalla qualità della progettazione didattica e dalla capacità di integrarla in modo consapevole nei processi educativi.

Ciò implica che il feedback mediato dall'IA debba essere progettato non solo in termini di contenuto, ma anche in termini di interazione. In particolare, diventa centrale la costruzione di situazioni didattiche in cui lo studente sia chiamato non soltanto a ricevere il feedback, ma a interpretarlo, discuterlo e rielaborarlo. In questo modo, il feedback può diventare esso stesso oggetto di apprendimento, favorendo lo sviluppo di consapevolezza rispetto ai criteri di validità matematica e alle forme di argomentazione richieste.

9. Conclusioni

Il contributo ha analizzato il ruolo del feedback automatizzato in matematica nel quadro del progetto PRIN "AI&F", mettendo in evidenza come questo ambito non rappresenti un semplice terreno di applicazione delle tecnologie di intelligenza artificiale, ma un contesto che ne problematizza in profondità le assunzioni di base. La natura epistemologica della matematica, caratterizzata da pluralità di strategie risolutive, centralità dell'argomentazione e complessità semiotica, impone infatti una riconfigurazione del concetto stesso di feedback, che non può essere ridotto a una restituzione di correttezza, ma deve essere inteso come processo di costruzione del significato.

In questa prospettiva, il contributo ha proposto una rilettura dei modelli di feedback alla luce della didattica della matematica, evidenziando la necessità di considerare simultaneamente dimensioni sintattiche, semantiche e argomentative. Tale integrazione si è rivelata centrale sia nella progettazione dei sistemi di feedback automatizzato, sia nell'interpretazione delle interazioni tra studente e intelligenza artificiale. In particolare, il modello AIM-FC ha consentito di descrivere il feedback come un ciclo iterativo e trasformativo, nel quale l'interazione con il sistema non esaurisce il processo, ma ne costituisce una fase intermedia, funzionale alla rielaborazione e alla revisione.

Le evidenze emerse dalla sperimentazione condotta dall'unità di Bari contribuiscono a chiarire ulteriormente il ruolo dell'intelligenza artificiale in questo contesto. Da un lato, l'interazione con sistemi generativi appare in grado di sostenere processi di esplicitazione e raffinamento argomentativo, favorendo una maggiore

consapevolezza da parte degli studenti rispetto alle proprie strategie e ai propri passaggi logici. Dall'altro lato, emergono limiti significativi legati all'affidabilità epistemica del feedback generato, soprattutto in presenza di errori concettuali, che richiedono un intervento critico e una mediazione da parte del docente.

Questi risultati suggeriscono che il valore dell'intelligenza artificiale nel feedback matematico non risiede nella possibilità di automatizzare la valutazione, ma nella capacità di ampliare e rendere più sostenibili i cicli di feedback, offrendo agli studenti ulteriori occasioni di confronto con la propria produzione. In tale prospettiva, l'IA può essere interpretata come un mediatore semiotico dinamico, capace di attivare processi di riflessione e rielaborazione, ma non come un sostituto del giudizio professionale del docente.

Dal punto di vista didattico, emerge con chiarezza che l'efficacia del feedback automatizzato dipende in modo cruciale dalla progettazione. La definizione dei compiti, la costruzione dei prompt, l'organizzazione delle fasi di lavoro e l'integrazione con momenti di discussione e istituzionalizzazione costituiscono elementi determinanti per orientare l'uso dell'IA in senso formativo. In assenza di una progettazione consapevole, il rischio è quello di ridurre il feedback a una restituzione superficiale o di introdurre forme di scaffolding eccessivamente direttive, che limitano l'autonomia dello studente. In conclusione, il contributo sostiene che lo sviluppo di sistemi di feedback automatizzato efficaci in matematica richiede un approccio disciplinarmente situato, capace di integrare tecnologia, epistemologia e pratiche di insegnamento. Più che interrogarsi sulle prestazioni dei sistemi di intelligenza artificiale in termini assoluti, appare necessario comprendere in quali condizioni tali sistemi possano essere utilizzati per sostenere processi di apprendimento significativi. In questa direzione, la sfida non è tanto tecnologica quanto didattica: progettare ambienti in cui l'interazione tra studente, docente e intelligenza artificiale contribuisca alla costruzione condivisa del significato matematico.

In questo senso, l'intelligenza artificiale non si limita a rendere più efficienti i processi esistenti, ma contribuisce a ridefinire le condizioni attraverso cui il feedback può diventare un dispositivo effettivamente formativo nei contesti della didattica della matematica.

Riferimenti bibliografici

Balacheff, N. (1999). Is argumentation an obstacle? *International Journal of Computers for Mathematical Learning, 4*(2-3), 153-176.

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for

a world with artificial intelligence. *British Journal of Educational Technology, 54*(5), 1160-1173. https://doi.org/10.1111/bjet.13337

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31. https://doi.org/-10.1007/s11092-008-9068-5

Brousseau, G. (1997). *Theory of didactical situations in mathematics*. Kluwer.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Chevallard, Y. (1999). L'analyse des pratiques enseignantes en théorie anthropologique du didactique. *Recherches en Didactique des Mathématiques, 19*(2), 221–266.

Corbin, T., Tai, J., & Flenady, G. (2025). Understanding the place and value of GenAI feedback: A recognition-based framework. *Assessment & Evaluation in Higher Education, 50*(5), 718-731. https://doi.org/10.1080/02602938.2025.2459641

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education, 162*, 104094. https://doi.org/-10.1016/j.compedu.2020.104094

Dubinsky, E., & McDonald, M. A. (2001). APOS: A constructivist theory of learning in undergraduate mathematics education research. In D. Holton (Ed.), *The teaching and learning of mathematics at university level* (pp. 275–282). Kluwer.

Duval, R. (2006). A cognitive analysis of problems of comprehension in a learning of mathematics. *Educational Studies in Mathematics, 61*(1-2), 103-131. https://doi.org-/10.1007/s10649-006-0400-z

Lipnevich, A. A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education, 6*, 720195. https://doi.org/10.3389/feduc.2021.720195

Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education, 35*(5), 501-517. https://doi.org/10.1080/02602931003786559

Pehkonen, E. (1997). The state-of-art in mathematical creativity. *ZDM – Mathematics Education, 29*(3), 63-67.

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35*(5), 535-550. https://doi.-org/10.1080/02602930903541015

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.

Sfard, A. (2008). *Thinking as communicating: Human development, the growth of discourses, and mathematizing*. Cambridge University Press.

Winstone, N. E., & Carless, D. (2019). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge.

II PARTE

Area tematica 1. IA e valutazione universitaria

II.1
Scrivere con l'IA generativa: usi, percezioni e nodi valutativi in ambito accademico

Angela Spinelli
Ricercatrice, Università di Roma Tor Vergata
angela.spinelli@uniroma2.it

Abstract

La diffusione dell'intelligenza artificiale nella didattica universitaria rende più complesso il rapporto tra scrittura, apprendimento e valutazione. L'articolo presenta uno studio di caso esplorativo nel corso di Didattica generale a.a. 2024/25 (Università di Roma Tor Vergata) i cui risultati mostrano un uso prevalentemente strumentale dei dispositivi basati su LLM, mentre restano marginali le pratiche metacognitive e valutative. Molti studenti si dichiarano disponibili ad un uso trasparente dell'IA, a condizione che le regole siano condivise ed esplicite. Dall'analisi emergono tre figure di *literacy* valutativa (strumentale, riflessiva e partecipativa), che offrono indicazioni per ripensare compiti e dispositivi in chiave più consapevole e responsabile.

Parole chiave: intelligenza artificiale; scrittura; valutazione; didattica universitaria; patto educativo.

The spread of artificial intelligence in university teaching makes the relationship between writing, learning and assessment increasingly complex. The article presents an exploratory case study conducted during the General Didactics course in the academic year 2024/25 (University of Rome Tor Vergata), the results of which show a predominantly instrumental use of Large Language Models, while metacognitive and evaluative practices remain marginal. Many students report being willing to use AI in a transparent and declared way, provided that rules are made explicit. The analysis identifies three forms of assessment literacy (instrumental, reflective and participatory), which offer indications for redesigning writing tasks and assessment devices in a more informed and responsible way.

Keywords: Artificial intelligence; Writing; Assessment; University teaching; Educational pact.

Introduzione

La diffusione dell'intelligenza artificiale (IA) nei contesti universitari ha reso meno nitido il rapporto tra scrittura, apprendimento e valutazione (García-Peñalvo, 2023; Hanemaayer, 2022). In poco tempo la produzione di testi è diventata il risultato di interazioni tra esseri umani e strumenti basati su Large Language Model (LLM), rendendo più incerti i confini tra ausilio, delega, sostituzione e sollecitando riflessioni di tipo normativo (UNESCO, 2019; Commissione europea, 2022). L'attuale fase di *postplagio* (Eaton, 2023; Claverini, 2024) è segnata, da un lato, dalla tendenza a rispondere con nuove forme di controllo e sorveglianza; dall'altro dall'urgenza di ripensare le condizioni educative della fiducia e della responsabilità autoriale (Floridi, 2022; Floridi, 2025; Jobin et al., 2019).

La ricerca internazionale si è concentrata soprattutto su tre assi: gli studi sull'*automated writing evaluation* e sugli effetti degli strumenti IA sulla qualità dei testi e sui processi di revisione (Krajka & Olszak, 2024; Pereira et al., 2024; Suwadi, 2023); le analisi, spesso allarmate, su plagio e tecnologie di *detection* (Eaton, 2023; Frye, 2023); i contributi che leggono l'IA come assistente alla scrittura per sostenere *brainstorming*, pianificazione e riscrittura (Pereira et al., 2024; Floridi, 2025). Parallelamente, nel campo della valutazione educativa si sono affermati i concetti di *assessment literacy* (Pastore, 2022) e, più recentemente, di *feedback literacy* (Nieminen & Carless, 2023), con una crescente attenzione alla capacità degli studenti di comprendere scopi della valutazione, interpretare criteri e usare il *feedback* per regolare l'apprendimento (Panciroli & Rivoltella, 2023; Ranieri et al., 2024).

Questi filoni, però, si incrociano solo in parte: l'uso dell'IA nella scrittura è raramente interrogato come pratica in cui si formano competenze valutative. L'IA è per lo più trattata come rischio o opportunità tecnica, mentre la *literacy* valutativa resta ancorata a scenari in cui il testo è prodotto dal solo soggetto umano e il *feedback* proviene da docenti o pari. Manca ancora una riflessione sistematica su come gli studenti leggano, negozino e interpretino la valutazione quando gli elaborati scritti sono il risultato di interazioni tra sé e la IA.

Questo contributo si colloca in questo spazio, proponendo di articolare il concetto di *literacy valutativa* degli studenti in ambiente IA-mediato a partire da uno studio di caso. L'obiettivo è, da un lato, ricostruire come gli studenti utilizzano l'IA nei compiti di scrittura e come si collocano rispetto a autenticità, trasparenza e dichiarazione d'uso; dall'altro, avviare una riflessione sulla *literacy* valutativa, intesa non come tratto individuale, ma come modo di posizionarsi dentro dispositivi valutativi in cui l'IA è parte dell'ecologia di apprendimento.

Il lavoro è guidato da tre domande: come gli studenti usano l'IA nei compiti di scrittura; come si collocano rispetto ad un suo uso trasparente e dichiarato; quali configurazioni discorsive emergono che possono essere lette come figure di *literacy* valutativa.

Metodologia

La ricerca si colloca all'interno di un più ampio progetto di analisi e riprogettazione delle attività del corso di Didattica generale (a.a. 2024/25, Università di Roma Tor Vergata) e adotta il disegno dello studio di caso esplorativo ad orientamento qualitativo (Yin, 2009; Trinchero, 2004). L'oggetto è l'uso che gli studenti di area pedagogica fanno degli strumenti basati su LLM nella scrittura accademica e nelle pratiche di apprendimento, con particolare attenzione alle dimensioni valutative. In continuità con il paradigma costruttivista, la ricerca assume come dati centrali le rappresentazioni che i soggetti producono del proprio agire e dei contesti in cui operano e mira a ricostruire i modelli che orientano le loro decisioni (Trinchero, 2004; Mortari & Ghirotto, 2019).

Il caso ha coinvolto oltre 350 studenti, in diversi momenti del semestre, attraverso una combinazione di strumenti quantitativi e qualitativi, nella logica dei *mixed methods* (Trinchero & Robasto, 2019). Sono stati utilizzati un questionario in ingresso sulle pratiche d'uso e le rappresentazioni dell'IA, un Forum dedicato nella piattaforma Moodle, sei focus group e un questionario in uscita specifico per la dichiarazione di utilizzo di IA, esteso anche ai laureandi.

	Numero	**Partecipanti**	**Rispondenti**	**Corpus**
Questionario in ingresso	1	284 (c.a.)	218	Quantitativo Qualitativo
Forum	1 29 argomenti	440	143 risposte 1967 visualizzazioni	Scritto
Focus group	6	23	23	Orale
Questionario uscita corsisti	1	440	296	Quantitativo Qualitativo
Questionario uscita tesisti	1	28	28	Quantitativo Qualitativo

Tab. 1 – Strumenti di rilevazione, partecipanti e corpus

Il questionario in ingresso (Likert + risposte aperte) rilevava familiarità e frequenza d'uso, funzioni attribuite e percezione di benefici/rischi (es. "Per quali attività usi l'IA?"; "Sei favorevole ad un uso didattico?"; "Sei disposto a dichiararne l'uso?"). Il Forum è stato strutturato in *thread* tematici e consegne progressive: gli studenti dovevano (i) condividere un'esperienza di co-scrittura umano–IA, (ii) discutere un caso di *bias* algoritmico e (iii) argomentare posizioni su autenticità, regole d'uso e ruolo docente, con interventi di sintesi/rilancio a presidio del dialogo. I focus group, semi-strutturati, esploravano pratiche quotidiane e criteri impliciti ("Quando usi l'IA nelle diverse fasi del compito?"; "Come decidi cosa è legittimo e cosa no?"; "Quali rischi temi? Quali opportunità?"). Il questionario in uscita riprendeva gli stessi assi, includendo una sezione di dichiarazione d'uso collegata al compito finale (strumenti usati, fasi, tipo di supporto, strategie di verifica e rielaborazione, esempi di prompt).

L'analisi dei risultati raccolti si è articolata su due livelli integrati. I dati dei questionari sono stati trattati mediante statistiche descrittive, con funzione di inquadramento del fenomeno. Il lavoro qualitativo sui *corpora* testuali è stato condotto seguendo il protocollo applicativo dell'analisi critica del discorso (Fabbro, 2024), in dialogo con l'impostazione costruttivista (Mortari, 2007) e con un approccio ispirato alla *grounded theory* (Tarozzi, 2008). A partire da una domanda guida ampia su usi, rappresentazioni e interazioni con l'IA, sono state definite quattro dimensioni di analisi (empirica, interpretativa, educativa, epistemologica/istituzionale), articolate in sotto-domande operative che hanno guidato la costruzione degli strumenti e la lettura dei testi.

Le categorie sono state organizzate in temi e sotto-temi, a partire da tre aree principali: IA come strumento di supporto all'attività cognitiva; come rischio per l'autenticità e il pensiero critico; come oggetto culturale ed educativo da mediare. L'intero processo analitico ha seguito un andamento spiraliforme: le categorie hanno orientato la lettura e la codifica, ma sono state rimesse in discussione alla luce delle ricorrenze e delle contraddizioni emerse, fino al raggiungimento di una saturazione interpretativa ritenuta soddisfacente (Mortari, Ghirotto, 2019; Salerni, 2002).

Risultati. Usi dell'IA nella scrittura e configurazioni di *literacy* valutativa

Nel contesto considerato, la scrittura accademica è al tempo stesso attività di apprendimento e dispositivo di valutazione. I dati mostrano che l'IA generativa si inserisce in questo intreccio con modalità differenziate, che riguardano sia le

pratiche d'uso sia il modo in cui gli studenti leggono il patto valutativo entro cui tali pratiche si collocano.

Le risposte ai questionari restituiscono un quadro in cui gli strumenti basati su LLM sono impiegati soprattutto come supporto operativo alla stesura dei testi: una parte significativa del campione dichiara di usarli, almeno occasionalmente, per cercare o riorganizzare informazioni e riferimenti, riformulare passaggi, migliorare la correttezza linguistica, sintetizzare contenuti. Sono molto meno frequenti gli usi che coinvolgono la dimensione metacognitiva, come chiedere alternative argomentative, simulare obiezioni, confrontare versioni del testo in funzione dei criteri, monitorare la coerenza complessiva. Queste pratiche sono presenti, ma minoritarie rispetto a un impiego orientato all'efficienza: il caso mostra una *literacy* digitale orientata da un atteggiamento strumentale e solo occasionalmente tradotta in pratiche dialogiche di autovalutazione.

Molti studenti esprimono accordo con l'idea che l'IA possa essere utilizzata in modo trasparente e dichiarato, a patto che siano esplicitate in modo comprensibile le regole: che cosa è lecito fare, che cosa va reso visibile, come si tiene insieme l'apporto dell'IA con la responsabilità personale rispetto al testo consegnato.

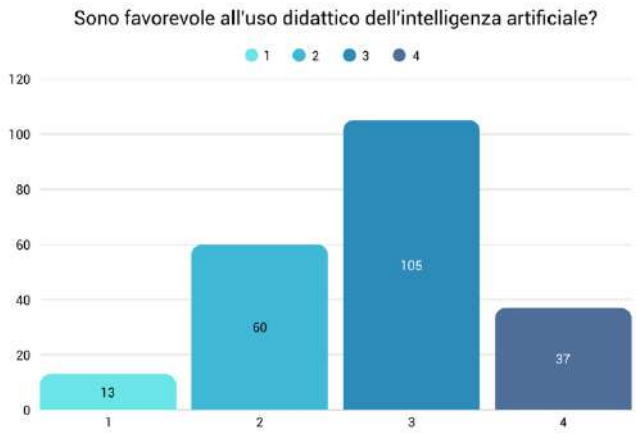


Fig. 1 – Favorevoli all'uso didattico della IA

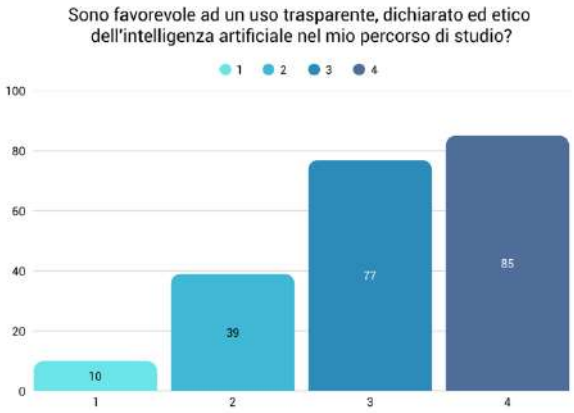


Fig. 2 – Favorevoli a dichiarare l'uso della IA

In modo coerente, una quota ampia del campione afferma che sarebbe disposta a dichiarare l'uso dell'IA nella stesura dei testi, se il contesto lo prevedesse e offrisse spazi e modalità per farlo; solo una minoranza si oppone, mentre una parte mantiene una posizione incerta, spesso legata al timore di conseguenze sul giudizio o alla percezione di criteri instabili.

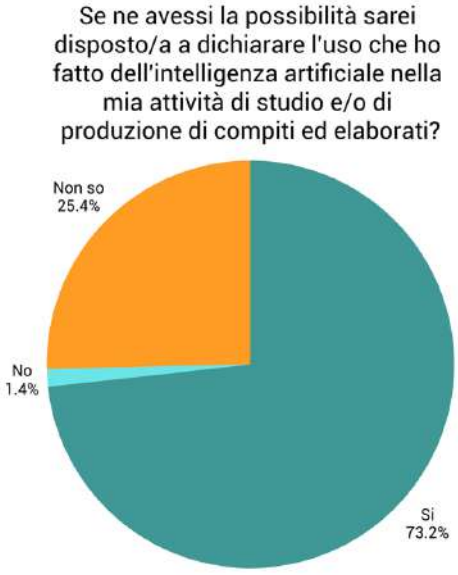


Fig. 3 – Favorevoli a dichiarare le modalità d'uso

Le discussioni nel Forum e nei focus group approfondiscono questa tensione. L'IA è riconosciuta come alleato nel gestire carichi di lavoro elevati, superare il blocco di scrittura, dare forma alle idee, ma emergono dubbi sulla giustizia del giudizio quando l'uso è diffuso ma disomogeneo, non regolato e difficilmente verificabile: chi non la usa teme di essere svantaggiato, chi la usa teme di essere sanzionato. In questo spazio ambiguo, disponibilità alla trasparenza e timore di "esporsi" coesistono e contribuiscono a definire la qualità della *literacy* valutativa.

Di seguito si riportano le principali configurazioni d'uso, individuate in base ai pattern discorsivi emergenti ricostruiti con l'analisi interpretativa a partire dalle combinazioni più frequenti e significative di codici tra categorie, temi e sotto-temi.

- IA come supporto: *literacy* valutativa strumentale. L'IA è un aiuto per "fare meglio il compito": organizzare materiali, chiarire passaggi, rifinire il testo. Il soggetto si interpreta come scrivente che mantiene il controllo delegando all'IA aspetti tecnici; sul piano valutativo l'attenzione resta sull'esito più che su criteri e obiettivi, che rimangono impliciti. La *literacy* valutativa assume qui una forma prevalentemente strumentale, orientata a soddisfare le attese percepite.
- IA come rischio per autenticità e pensiero critico: *literacy* valutativa riflessiva. L'IA è tematizzata come minaccia: appiattimento dello stile, perdita della "voce" personale, indebolimento dell'impegno cognitivo. Il soggetto assume una postura critica, interrogando il confine tra aiuto legittimo e delega eccessiva e chiedendosi fino a che punto il testo possa dirsi proprio. La *literacy* valutativa assume una forma riflessiva: gli studenti valutano l'impatto dell'uso dell'IA rispetto a obiettivi formativi, criteri di autenticità e senso della valutazione.
- IA come oggetto culturale ed educativo da mediare, *literacy* valutativa partecipativa. Il fuoco si sposta sul piano istituzionale: l'IA è discussa in termini di regole, equità, diritti e doveri, chiedendosi chi decide le norme e come si assicura una valutazione giusta in presenza di usi diseguali. Il soggetto si posiziona come cittadino, chiedendo linee guida comprensibili e coerenza tra discorso ufficiale e prassi. La *literacy* valutativa assume un profilo partecipativo: non solo decodifica criteri, ma rivendica la possibilità di intervenire sui dispositivi di valutazione e sulle condizioni di trasparenza e giustizia.

Nel complesso, la *literacy* valutativa nell'interazione con l'IA appare un insieme di posizionamenti che oscillano tra questi tre poli. La prevalenza di usi

strumentali non esaurisce il quadro: accanto ad essa si intravedono risorse riflessive e partecipative che, se riconosciute e coltivate, possono diventare punto di partenza per riprogettare compiti, criteri e dispositivi di *feedback*.

Discussione

La maggior parte degli studenti descrive gli strumenti basati su LLM come utili per cercare e riorganizzare informazioni, sistemare la forma del testo, sintetizzare materiali; solo una minoranza riferisce usi più metacognitivi, come confrontare versioni, simulare obiezioni, verificare la coerenza rispetto ai criteri. Il caso conferma quanto emerso in altri studi, che segnalano un forte orientamento alla semplificazione del lavoro e un coinvolgimento più debole sul piano della revisione concettuale (Krajka & Olszak, 2024; Pereira et al., 2024; Suwadi, 2023). L'elemento specifico di questo studio è la lettura in chiave valutativa: l'IA è impiegata per "*fare meglio il compito*", ma raramente per interrogarsi sul significato del "*fare bene*" rispetto a criteri e obiettivi formativi.

Per quanto riguarda l'uso trasparente dell'IA, i dati mostrano che molti studenti sarebbero disposti a dichiararlo, a condizione che esistano regole chiare e condivise. L'idea di un uso trasparente ed eticamente regolato incontra un consenso ampio, ma resta sospesa finché le condizioni istituzionali non sono esplicitate. In relazione al plagio, la questione non è solo se l'IA favorisca pratiche scorrette, ma come l'istituzione costruisca un quadro di fiducia in cui sia possibile renderne visibile il contributo senza trasformarlo automaticamente in sospetto. In questa prospettiva la *literacy* valutativa non è solo un insieme di abilità individuali, ma dipende da come l'università rende leggibile e discutibile il patto che regola la valutazione in presenza dell'IA.

Rilette con gli occhiali della valutazione, le tre configurazioni d'uso descritte precedentemente rimandano ad altrettante figure. Il collegamento tra configurazioni e dimensione valutativa emerge perché, nei dati, l'IA non è interpretata solo come uno strumento di scrittura, piuttosto come un dispositivo che esplicita il rapporto tra attese, criteri e regole della prova in un contesto che coinvolge anche la dimensione istituzionale, oltre che quella pedagogica. Le posture valutative possono essere quindi sintetizzate in:

- strumentale: IA serve a soddisfare attese date per scontate;
- critico-riflessiva: l'uso è occasione per interrogare criteri e senso della prova;
- partecipativa: centratura sulle regole, sulla coerenza delle policy e sulla possibilità di negoziare dispositivi valutativi.

La letteratura su *assessment* e *feedback literacy* ha finora lavorato soprattutto in contesti in cui il testo è prodotto dal solo soggetto umano; in questo caso, la presenza dell'IA introduce una nuova configurazione che gli studenti devono imparare a considerare. Il quadro conferma tendenze già note - uso orientato all'efficienza, preoccupazioni più legate alla perdita di autenticità e di voce che al plagio in senso tradizionale, percezione ambivalente dell'IA come risorsa e rischio - ma sposta l'attenzione su come gli studenti interpretano la valutazione quando il testo è il risultato di un intreccio di azioni umane e algoritmi. Le tre figure vanno intese come modi di posizionarsi rispetto a compiti, strumenti e regole; il fatto che coesistano posizionamenti strumentali, riflessivi e partecipativi indica che il margine di trasformazione dipende dalle scelte didattiche e valutative che si introducono.

I limiti dello studio risiedono nel fatto di essere un caso singolo, molto situato, difficilmente trasferibile senza una lettura contestuale. I dati quantitativi sono auto-dichiarati e sono stati utilizzati con funzione descrittiva, rinunciando intenzionalmente a sfruttare il potenziale inferenziale del *dataset*, che sarà oggetto di elaborazioni successive.

A partire da questi elementi, le linee di lavoro future riguardano la progettazione di compiti di scrittura e dispositivi valutativi che rendano visibile e discutibile l'uso dell'IA (ad esempio attraverso sezioni di dichiarazione d'uso e rubriche che considerano il processo oltre al prodotto), la messa a punto del modello di *literacy* valutativa IA-mediata, e l'inclusione del punto di vista di docenti e istituzioni, costantemente evocato dagli studenti ma non indagato in questo studio. Comprendere come i diversi attori definiscono e vivono il patto valutativo in presenza dell'IA è il passo necessario per passare dalla descrizione delle tensioni alla sperimentazione di pratiche che le affrontino in modo esplicito.

Conclusione

Lo studio che abbiamo presentato mostra che l'uso dell'IA nella scrittura accademica non riguarda solo nuovi strumenti, ma il modo in cui gli studenti si collocano dentro la valutazione. Le pratiche descritte convivono con domande esplicite su autenticità, giustizia del giudizio, coerenza delle regole. In questo senso, la *literacy* valutativa appare come un nodo educativo rilevante: non è un semplice complemento tecnico, ma un punto di passaggio tra apprendimento, responsabilità autoriale e fiducia nei dispositivi di valutazione.

Ne derivano alcune indicazioni operative per i docenti: 1) esplicitare aspettative e criteri distinguendo ciò che conta nel prodotto e ciò che conta nel pro-

cesso; 2) definire regole condivise di uso e dichiarazione dell'IA, considerandole come parte integrante del patto educativo; 3) progettare consegne che rendano osservabile il percorso (bozze, revisioni, diario d'uso) e riducano l'uso strumentale; 4) promuovere una *prompt literacy* critica che includa la capacità di interrogare, verificare, confrontare, mantenendo il controllo su errori, *bias* e allucinazioni; 5) riposizionare la valutazione in chiave formativa durante tutto il percorso di apprendimento, con momenti di metariflessione e *feedback* che sostengano un uso consapevole ed etico dell'intelligenza artificiale.

Riferimenti bibliografici

Claverini, C. (2024). Etica della fiducia e intelligenza artificiale generativa nell'epoca del post plagio. *Mizar. Costellazioni di pensieri,* 21, 103-114.

Commissione europea. (2022). *Orientamenti etici per gli educatori sull'uso dell'intelligenza artificiale e dei dati nell'insegnamento e nell'apprendimento*. Ufficio delle pubblicazioni dell'Unione europea.

Eaton, S. E. (2023). Postplagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. *International Journal for Educational Integrity, 19*(1), 1-10.

Fabbro, F. (2024). *L'analisi critica del discorso per la ricerca educativa. Teorie, metodi, strumenti*. Roma: Carocci.

Floridi, L. (2022). *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*. Milano: Raffaello Cortina.

Floridi, L. (2025). *Distant writing: Literary production in the age of artificial intelligence*. Vienna: Centre for Digital Ethics (CEDE).

García-Peñalvo, F. J. (2023). The perception on artificial intelligence in higher education: Opportunity, disruption or panic? *Education in the Knowledge Society,* 24, 1-9.

Hanemaayer, A. (Ed.). (2022). *Artificial intelligence and its discontents: Critiques from the social sciences and humanities*. Cham: Springer.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence,* 1, 389-399.

Krajka, J., & Olszak, I. (2024). Artificial intelligence tools in academic writing instruction: Exploring the potential of on-demand AI assistance in the writing process. *Roczniki Humanistyczne, 72*(6), 123-140.

Mortari, L. (2007). *Cultura della ricerca pedagogica*. Roma: Carocci.

Mortari, L., Ghirotto L. (2019). *Metodi per la ricerca educativa*. Roma: Carocci.

Nieminen, J. H., & Carless D. (2023). Feedback literacy: a critical review of an emerging concept. *Higher Education. The International Journal oh Higher Education Research,* 85, 1381-1400.

Panciroli, C., & Rivoltella, P. C. (2023). *Pedagogia algoritmica. Per una riflessione educativa sull'intelligenza artificiale*. Brescia: Scholé.
Pastore, S. (2022). Assessment literacy in the higher education context: A systematic review. *Intersection: A journal at the intersection of assessment and learning*, 4(1).
Pereir, R., Reis, I. W., Ulbricht, V., & dos Santos N. (2024). Generative artificial intelligence and academic writing: An analysis of the perceptions of researchers in training. *Management Research: Journal of the Iberoamerican Academy of Management, 22*(4), 429-450.
Ranieri, M., Cuomo, S., & Biagini, G. (2024). *Scuola e intelligenza artificiale. Percorsi di alfabetizzazione critica*. Roma: Carocci.
Salerni, A. (2002). *Tecniche e strumenti di rilevazione*. In P. Lucisano; A. Salerni (a c. di), *Metodologia della ricerca in educazione e formazione*. Roma: Carocci, 149-287.
Suwadi, S. (2023). A utilization of artificial intelligence in learning activities in higher education. *EDUTEC: Journal of Education and Technology, 7*(2), 596-604.
Tarozzi, M. (2008). *Che cos'è la grounded theory*. Roma: Carocci.
Trinchero, R. (2004). *I metodi della ricerca educativa.* Roma-Bari: Laterza.
Trinchero, R., & Robasto D. (2019). *I mixed methods nella ricerca educativa*. Milano: Mondadori.
UNESCO. (2019). *International Conference on Artificial Intelligence and Education: Planning education in the AI era. Lead the leap. Final report*. UNESCO.
Yin, R. K. (2009). *Case study research: Design and methods*. SAGE.

II.2
Integrare l'IA nell'assessment: il modello rAIse

Nadia Sansone
Prof.ssa Ordinaria in PAED-02/B, UnitelmaSapienza Università di Roma
nadia.sansone@unitelmasapienza.it
Ilaria Bortolotti
PhD, Cultrice della Materia e Assistente alla cattedra in PAED-02/B, UnitelmaSapienza Università di Roma
ilaria.bortolotti@unitelmasapienza.it

Abstract

Si presenta uno studio esplorativo con 17 studenti universitari impegnati in un'e-tivity che integra GenAI conversazionale, Assessment as Learning e il modello rAIse come scaffold procedurale. Il corpus comprende log conversazionali (N=68), questionari rAIse pre/post e peer feedback strutturato. I log sono stati analizzati tramite un framework di codifica basato su indicatori ordinali di funzione cognitiva attribuita all'AI e livello di iterazione, integrati da marker di regolazione (meta-prompting, critica e rifiuto dell'output). I risultati evidenziano, in una quota del campione, l'emergere di strategie metacognitive di controllo dell'interazione; la triangolazione tra profili comportamentali, autovalutazione e peer feedback rivela una relazione tra profilo interattivo e allineamento tra competenza percepita e agita, con il feedback dei pari come fattore correttivo nei profili più superficiali. Si discutono implicazioni progettuali per attività che trasformano la conversazione con l'AI in pratica valutativa e autoregolativa.

Parole chiave: IA conversazionale; Valutazione formativa; rAIse; Assessment as learning; Etica dell'IA.

Abstract
This paper presents an exploratory study with 17 university students engaged in an e-tivity integrating conversational GenAI, Assessment as Learning, and the rAIse model as a procedural scaffold. The dataset includes chat logs (N=68), rAIse pre/post questionnaires, and structured peer feedback. Logs were analyzed using a coding framework based on ordinal indicators of cognitive function attributed to AI and iteration level, supplemented by regulation markers (meta-prompting, output critique and rejection). Findings show that a subset of students developed metacognitive strategies for governing AI interaction. Triangulation across behavioral profiles, self-assessment, and peer feedback reveals a relationship between interaction profile and alignment between perceived and enacted competence, with peer feedback acting as a corrective mechanism in

more superficial profiles. Design implications are discussed for activities that transform AI conversation into a formative and self-regulatory practice.

Keywords: Conversational AI; Formative assessment; rAIse; Assessment as learning; AI ethics.

1. Generative AI, dialogo metacognitivo e valutazione per apprendere

L'adozione della Generative AI (GenAI) nei contesti educativi apre possibilità rilevanti, ma introduce anche rischi che rendono necessario spostare l'attenzione dal *"cosa produce"* l'AI al *"come si interagisce"* con l'AI. Da un lato, infatti, i Large Language Models (LLM) possono sostenere studio e scrittura offrendo spiegazioni, riformulazioni, esempi e ipotesi di soluzione, oltre a uno spazio di dialogo che facilita chiarificazione e rielaborazione. Dall'altro, il loro uso comporta criticità note: inaccuratezze e allucinazioni, opacità delle fonti, bias, risposte persuasive anche quando non fondate, e il rischio di uso sostitutivo (delega cognitiva) che riduce l'impegno riflessivo. In ambito didattico, queste contrapposizioni fanno emergere una domanda: in che modo l'interazione con la GenAI può trasformarsi in un percorso di apprendimento, preservando la capacità dello studente di valutare, verificare e decidere in prima persona? A tal fine, occorre spostare il focus sulla conversazione come spazio di costruzione del significato. La formulazione e riformulazione delle richieste (i prompt) vanno quindi letti come momenti in cui il pensiero viene esternalizzato e reso regolabile (Vygotskij, 1978) e la conoscenza come esito che emerge nello spazio tra domanda e risposta, nella gestione di alternative, obiezioni e chiarimenti (Bakhtin, 1981). Questa dinamica richiede tuttavia un presidio metacognitivo, inteso come conoscenza e controllo dei propri processi cognitivi, attraverso pianificazione, monitoraggio e regolazione delle strategie (Flavell, 1979). Ne consegue che l'interazione con la GenAI è educativa non "di per sé", ma nella misura in cui lo studente governa la conversazione: chiarisce obiettivi, valuta criticamente l'output, decide revisioni e verifica la tenuta delle informazioni.

Studi recenti evidenziano che l'integrazione della GenAI nei contesti educativi richiede un ripensamento delle pratiche valutative tradizionali (Bearman et al., 2023), poiché gli strumenti generativi alterano le condizioni di autenticità e responsabilità del processo di apprendimento. Parallelamente, la questione dell'AI literacy emerge come competenza trasversale irrinunciabile: essa non si riduce all'uso tecnico degli strumenti, ma include la capacità critica di interpretare, negoziare e governare le risposte del sistema (Long, Magerko, 2020; Ng et

al., 2021). In questo scenario, l'uso di LLM come partner dialogici per la valutazione formativa rappresenta un terreno di ricerca emergente (Kasneci et al., 2023), in cui la qualità dell'interazione - più che la qualità dell'output -diventa l'indicatore centrale dell'apprendimento (Mollick & Mollick, 2023).

Su queste basi, l'attività didattica qui descritta mira a collegare AI literacy dialogica e valutazione per apprendere, attraverso il paradigma dell'*Assessment as Learning* (AaL) per cui la valutazione è una parte fondamentale del processo di apprendimento: lo studente impara a valutare il proprio lavoro mentre lo svolge, usando criteri chiari per osservare ciò che sta facendo e decidere come migliorare o correggere la rotta (Earl, 2003). In questo modo attiva cicli di autoregolazione: pianifica, mette in atto, rivede criticamente e adatta strategie e decisioni (Zimmerman, 2002). Perché ciò avvenga, l'autovalutazione deve essere sostenuta da criteri trasparenti (Sadler, 1989) e da feedback strutturati: rubriche e dispositivi di confronto aiutano a trasformare l'interazione con l'AI da richiesta di esecuzione a pratica di discernimento (Nicol & Macfarlane-Dick, 2006), riducendo sia giudizi impressionistici sia delega epistemica. In questa cornice, il modello rAIse (Sansone, submitted) articola l'interazione con la GenAI in cinque fasi sequenziali e ricorsive, progettate per rendere osservabili e allenabili i processi metacognitivi sottostanti. La fase Reflect orienta lo studente alla definizione esplicita dell'obiettivo, del contesto e delle conoscenze pregresse prima di formulare qualsiasi richiesta: si tratta di un momento di attivazione della pianificazione intenzionale che impedisce l'accesso impulsivo allo strumento. La fase Ask riguarda la costruzione del prompt come atto linguistico consapevole e vincolato: la richiesta non è generica ma ancorata a un obiettivo dichiarato, a un formato atteso e a criteri espliciti di qualità, trasformando la scrittura del prompt in un'operazione cognitiva complessa analoga alla formulazione di un'ipotesi di lavoro. Nella fase Interpret lo studente analizza criticamente l'output ricevuto, verificandone coerenza interna, completezza e adeguatezza rispetto all'obiettivo iniziale: è qui che si attiva il confronto tra quanto atteso e quanto prodotto, condizione necessaria per qualsiasi regolazione (Nicol & Macfarlane-Dick, 2006). La fase Source introduce la valutazione della fondatezza e della provenienza delle informazioni contenute nella risposta del sistema, contrastando attivamente la tendenza alla delega epistemica che caratterizza gli usi più superficiali della GenAI. La fase Evolve, infine, guida la revisione dell'elaborato e la riflessione sul processo complessivo – cosa ha funzionato nell'interazione, cosa richiederebbe un'iterazione diversa – chiudendo il ciclo metacognitivo e aprendo a un uso progressivamente più consapevole dello strumento. Questo schema ha una duplice funzione: da un lato fornisce allo studente un protocollo di scaffolding che struttura la conversazione come sequenza di

operazioni cognitive intenzionali (Vygotskij, 1978); dall'altro costituisce per il ricercatore un framework di codifica che rende l'interazione con l'AI analizzabile in termini di qualità del prompt, profondità della revisione e attribuzione di funzione cognitiva al sistema. In questo senso rAIse non è solo un dispositivo didattico, ma uno strumento di ricerca educativa situata (Sansone, submitted).

2. Metodologia della Ricerca

2.1 Obiettivi e Domande di Ricerca

Lo studio adotta un approccio mixed-methods di tipo esplorativo, finalizzato a indagare le dinamiche di interazione tra studenti universitari e LLM in contesti di apprendimento formale.

Lo studio è guidato da due domande di ricerca (Research Questions - RQ):

– RQ1 (Sviluppo Metacognitivo): In che misura emergono strategie di regolazione dell'interazione, soprattutto nelle fasi di autovalutazione e revisione?
– RQ2 (Profili interattivi e Allineamento percettivo): Quali profili di interazione con i LLM emergono dai log conversazionali e in che misura tali profili si associano al grado di allineamento tra competenza percepita e competenza agita alla luce del peer feedback?

2.2 Partecipanti e Procedura

La ricerca è stata condotta nell'ambito dell'insegnamento "Tecnologie per l'Apprendimento" (CdL online in Scienze e Tecniche Psicologiche, II anno), attraverso l'e-tivity "Imparare ad imparare con l'AI", cui hanno partecipato 42 studenti. Ai fini dell'analisi, il campione (N=17) include coloro che hanno completato le 3 fasi previste nel periodo 21 ottobre–22 novembre 2025:

1. rAIse. Nel webinar di lancio, viene introdotto e sperimentato in diretta il modello rAIse come scaffold conversazionale e protocollo di lavoro che gli studenti sono chiamati a riutilizzare nelle attività successive. Il modello guida l'interazione con il LLM [1] attraverso una sequenza di stimoli: definizione

1 Agli studenti è stata offerta autonomia nella selezione dello strumento; sebbene ChatGPT sia risultato prevalente, nel corpus sono presenti anche interazioni condotte con Gemini, Claude e Perplexity.

dell'obiettivo, formulazione di richieste controllate, interpretazione critica dell'output, verifica/triangolazione, sintesi e miglioramento.

2. Auto-Valutazione. Nel secondo webinar si avvia il blocco di autovalutazione: (a) gli studenti chiedono all'AI di valutare un proprio elaborato con una richiesta formulata "come farebbero normalmente", che costituisce la baseline per l'analisi del prompting. Successivamente, (b) riformulano la strategia applicando il rAIse e introducendo criteri e pratiche di regolazione (es. metaprompting), (c) co-costruiscono con l'AI una rubrica (Panadero e Jönsson, 2013), la applicano all'elaborato e la raffinano attraverso iterazioni motivate, documentando le revisioni [2].
3. Nel terzo webinar si svolge un peer feedback strutturato secondo una griglia di analisi centrata sul processo, derivata dal framework rAIse: ogni studente ha esaminato la conversazione anonimizzata di un collega, valutando la qualità dei prompt prodotti, la funzione cognitiva attribuita all'AI e la presenza di strategie di regolazione dell'interazione. Il feedback è stato espresso in forma scritta attraverso un modulo digitale, con commenti discorsivi su punti di forza e criticità del processo osservato. Il Webinar ha costituito il momento di confronto diretto: i feedback sono stati restituiti collettivamente, consentendo agli studenti di discutere le osservazioni ricevute, confrontare strategie diverse e riflettere sullo scarto tra competenza percepita e competenza agita nel proprio processo. L'anonimizzazione delle conversazioni condivise ha garantito un clima di analisi critica orientata al miglioramento piuttosto che alla valutazione della persona.

2.3 Strumenti di Analisi

Per rispondere alle Research Question (RQ), l'analisi ha combinato indicatori comportamentali ricavati dai log conversazionali con misure percettive raccolte tramite moduli Google. A fini descrittivi, i log (N=68) sono stati preliminarmente caratterizzati in termini di qualità strutturale del prompt (Quali, 0-3) e lunghezza media degli interventi (L_media, espressa in numero di caratteri per turno), trattati come indicatori di base del livello di contestualizzazione dell'interazione.

– RQ1: l'analisi longitudinale dei log si è focalizzata sulle fasi di revisione e au-

2 La conversazione di co-costruzione e applicazione della rubrica si svolge in modo asincrono tra il secondo e terzo webinar.

tovalutazione per identificare la funzione cognitiva implicitamente attribuita all'AI (Func, 0–3: da uso esecutivo-sostitutivo a uso critico-regolativo), il livello di iterazione trasformativa (Iter, 0-3) e la presenza di marker di regolazione, tra cui meta-prompting, rifiuto o critica motivata dell'output, e richieste di validazione del prompt prima della risposta.
– RQ2: i profili interattivi sono stati derivati dalla combinazione di Quali iniziale, Func, Iter e marker di regolazione. L'allineamento percettivo è stato stimato confrontando la competenza agita (log) con quella percepita, rilevata tramite questionario rAIse pre/post (N=17) su quattro dimensioni di AI literacy e il peer feedback strutturato (N=17), raccolto in forma scritta tramite modulo digitale e discusso nel webinar conclusivo.

Il questionario rAIse è uno strumento costruito ad hoc, somministrato in forma digitale in modalità pre/post, articolato in tre blocchi. Il primo blocco (pre) rileva la percezione di AI literacy su quattro dimensioni tramite scala Likert a cinque livelli: competenza nell'uso dell'AI per ottenere informazioni, competenza per il supporto professionale, capacità di analisi critica degli output, consapevolezza delle implicazioni etiche (privacy, bias, allucinazioni). Il secondo blocco, compilato durante l'interazione, è strutturato in corrispondenza delle cinque fasi del modello rAIse: indaga l'obiettivo della conversazione e il livello di preparazione sull'argomento (Reflect), la chiarezza e l'efficacia del prompt formulato (Ask), le modalità di modifica del prompt e la valutazione critica della risposta ricevuta – inclusa la presenza percepita di bias o imprecisioni (Interpret), il ricorso a fonti aggiuntive per verifica e approfondimento (Source), la riflessione su quanto appreso sul proprio processo e le intenzioni di miglioramento (Evolve). Il terzo blocco (post) ripropone le stesse quattro dimensioni di AI literacy del blocco iniziale - consentendo il confronto pre/post - e aggiunge un item sulla facilità d'uso del modello rAIse come scaffold. Nel complesso, il questionario permette di triangolare la competenza percepita con quella agita (rilevata dai log), costituendo la misura di allineamento percettivo utilizzata per RQ2.

3. Risultati

I risultati sono illustrati per domanda di ricerca.

3.1 RQ1: Sviluppo metacognitivo

L'analisi degli indicatori Func e Iter, integrata dai marker di regolazione, evidenzia l'emergere, in una parte consistente del campione, di strategie di controllo metacognitivo dell'interazione che vanno oltre la semplice richiesta di output. Si distinguono tre modalità ricorrenti, non mutuamente esclusive.

La prima è la regolazione per chiarezza: lo studente interviene attivamente per assicurare usabilità e precisione dell'output, contestando descrittori vaghi o sovrapposizioni tra criteri ("Per definire ogni criterio, potresti usare un termine unico. Le frasi sono troppo lunghe e vaghe, dammi 2 opzioni per ogni livello"). La seconda è la regolazione per controllo, riconducibile al meta-prompting: circa un terzo del campione richiede all'AI di valutare il prompt prima di rispondere ("Prima di rispondere, dimmi se il prompt che ho formulato è completo; se non lo è, dimmi come migliorarlo"), invertendo il flusso dialogico standard e rendendo visibile un presidio metacognitivo sulla qualità della propria richiesta. La terza è la regolazione per raffinamento: lo studente gestisce attivamente la dissonanza tra la propria intenzione e la risposta ricevuta, rifiutando proposte compiacenti e negoziando il contenuto ("Non sono molto d'accordo sulle esemplificazioni che fai... Vorrei sostituire il criterio 3 con un altro criterio: con cosa proponi di sostituirlo?"). Queste strategie si associano a valori più elevati di Func e Iter e compaiono prevalentemente nelle fasi di co-costruzione della rubrica e di revisione dell'elaborato, dove il compito richiede esplicitamente di confrontare un output con criteri dichiarati.

3.2 RQ2: Profili comportamentali e Alignment percettivo

La tabella 1 sintetizza tipologia e diffusione dei profili comportamentali rintracciati e rispettivi livelli di allineamento percettivo.

Profilo	%	**Caratteristiche chiave (comportamento in chat)**	**Allineamento percettivo**
Apprendente Riflessivo	29%	Meta-prompting; negoziazione del senso; ricerca di feedback critico/severo; iterazioni trasformative	Alto: autovalutazione coerente con le strategie osservate e con il peer feedback
Architetto Strategico	18%	Prompt strutturati (ruolo/obiettivo/vincoli/formato); controllo del contesto; orientamento all'efficienza	Alto: consapevolezza adeguata; feedback dei pari coerente
Orientato al compito	18%	Iterazioni frequenti ma prevalentemente tecniche/formali; AI come tutor per ottimizzare il prodotto	Medio: attenzione più al prodotto che al processo; coerenza parziale tra percepito e agito
Pragmatico	23%	Buon avvio, bassa persistenza; chiusura al "primo risultato utile"; uso utilitaristico	Medio: consapevolezza pragmatica; allineamento variabile
Apprendente Superficiale	12%	Prompt direttivi e brevi; accettazione passiva; iterazione minima o assente	Basso: eccessiva sicurezza, discrepanza tra percepito e agito - spesso corretta dal peer feedback

Tab. 1 – Profili di AI Literacy rilevati

I profili si distribuiscono lungo un continuum che va dall'uso delegante a quello critico-regolativo. I profili riflessivi (Apprendente Riflessivo 29%; Architetto Strategico 18%) si caratterizzano per l'uso intenzionale di vincoli, ruoli e criteri nel prompt, per iterazioni trasformative piuttosto che ripetitive, e per una ricerca attiva di feedback critico anche quando l'output appare soddisfacente. In questi profili l'autovalutazione risulta generalmente coerente con le strategie osservate nei log, indicando una consapevolezza più centrata sul processo che sul prodotto. I profili intermedi (Orientato al Compito 18%; Pragmatico 23%) mostrano un avvio controllato, ma una chiusura precoce al primo risultato utile, con iterazioni frequenti ma prevalentemente formali. I profili superficiali (Apprendente Superficiale 12%) si caratterizzano per prompt direttivi e brevi, accettazione passiva dell'output e iterazione assente o minima.

L'analisi dell'allineamento percettivo – stimato confrontando il Delta pre/post del questionario rAIse con la valutazione dei pari – rivela un pattern rilevante: nei profili superficiali emerge sistematicamente sovrastima, con studenti che dichiarano elevata efficacia pur in presenza di interazioni valutate dai pari come brevi e poco controllate ("Prompt poco chiari, corti e freddi… manca la

voglia di chiarire meglio quello che si vuole"). In questi casi il peer feedback ha operato come dispositivo correttivo, rendendo visibile uno scarto tra percezione e processo che la sola autovalutazione non avrebbe intercettato. Nei profili più riflessivi, viceversa, il delta pre/post del questionario rAIse risulta coerente con le strategie osservate nei log, confermando una consapevolezza di processo consolidata e non meramente dichiarativa.

4. Discussione

L'emergere di meta-prompting, critica dell'output e iterazioni trasformative segnala che una parte del campione utilizza l'AI come supporto alla regolazione del pensiero piuttosto che come strumento esecutivo: "va bene, ti faccio notare però che alcuni dei punti sono già presenti nella relazione: per esempio, titoli o sottotitoli esplicativi, approfondimenti espliciti, citazioni di momenti specifici ed esempi concreti, risposte analitiche". In termini metacognitivi, i log mostrano passaggi riconducibili a pianificazione, monitoraggio e regolazione delle strategie (Flavell, 1979), nonché una gestione più consapevole della "polifonia" informativa: la conversazione diventa uno spazio in cui vengono messe in gioco alternative, obiezioni e chiarimenti, cioè una costruzione di significato nello scarto tra domanda e risposta (Bakhtin, 1981): "Ho appreso che l'uso di questi strumenti richiede molta consapevolezza e una partecipazione attiva: sia nello sviluppo delle competenze [...] sia nell'esercizio della propria volontà, imponendo all'AI una direzione definita da noi e non il contrario. Queste sono le basi imprescindibili per costruire un dialogo efficace con i modelli e utilizzarli come strumenti di potenziamento delle proprie abilità e conoscenze, nello studio come nel lavoro". In questa prospettiva, l'interazione con il LLM risulta formativa quando lo studente mantiene l'iniziativa epistemica, trattando l'output come ipotesi da discutere e verificare, non come risposta autoritativa: "L'approccio iterativo è efficace: non ti fermi alla prima versione, ma chiedi raffinamenti, esempi multipli e specificazioni".

La triangolazione tra profili comportamentali (log), autovalutazione (questionari rAIse pre/post) e peer feedback evidenzia un pattern rilevante: i profili più riflessivi tendono a mostrare un miglior allineamento tra competenza percepita e competenza agita, mentre nei profili superficiali emerge più frequentemente sovrastima. Questo risultato è particolarmente significativo alla luce dei rischi della GenAI (risposte persuasive e apparentemente "buone" anche quando epistemicamente fragili), perché mostra che la consapevolezza non coincide automaticamente con l'efficacia del prodotto, ma dipende dalla qualità del processo

di controllo e verifica. In tal senso, il peer feedback agisce come dispositivo di trasparenza (Hattie & Timperley, 2007) del processo e sostiene una cultura del criterio, contribuendo a spostare l'attenzione dall'esito immediato al modo in cui l'esito è stato costruito.

5. Conclusioni: impatti, sviluppi futuri e condizioni di successo

Nel complesso, i risultati indicano che un percorso procedurale come quello proposto può trasformare l'uso della GenAI da produzione di output a processo governato di intenzionalità, controllo e revisione: "Il modello ha costituito un reminder strutturato di verifiche pre e post interazione, e mi ha aiutato a leggere criticamente le risposte". Il valore formativo non sta nel "prompt migliore" in sé, ma nel rendere la conversazione un dispositivo di metacognizione e valutazione per apprendere: quando lo studente lavora con criteri, verifica e iterazioni motivate, sviluppa giudizio e autoregolazione (Earl, 2003; Zimmerman, 2002) utile per l'apprendimento a lungo termine (Boud & Falchikov, 2007). Da qui derivano tre implicazioni: (1) Un'attività strutturata può sostenere una AI literacy dialogica e situata, osservabile nel passaggio dalla delega a pratiche di domanda, monitoraggio critico, verifica e revisione. (2) Quando i criteri sono parte integrante del compito, la GenAI può supportare forme più solide di autovalutazione, riducendo l'uso impressionistico e favorendo decisioni di miglioramento tracciabili. (3) Le tracce conversazionali permettono di leggere il processo e di confrontare competenza agita e percepita, rendendo visibili pattern di consapevolezza o sovrastima e rafforzando la feedback literacy (Carless & Boud, 2018).

Gli sviluppi futuri riguardano la validazione del framework di codifica, l'ampliamento del campione con repliche in altri contesti e un'analisi più fine delle traiettorie tra fasi, includendo anche misure di esito sul prodotto come variabile secondaria. Sul piano applicativo, emergono quattro condizioni di successo: criteri espliciti incorporati nel compito, verifica richiesta e valutata (validazione/triangolazione), feedback strutturato tra pari e uso dei log come supporto a riflessione e autovalutazione. In sintesi, la GenAI sostiene apprendimenti significativi quando è inserita in attività che obbligano a esplicitare obiettivi, usare criteri, controllare criticamente e verificare evidenze, mantenendo in mano allo studente la responsabilità epistemica.

Riferimenti bibliografici

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays*. (M. Holquist, Ed.; C. Emerson, M. Holquist, Trans.). Austin (TX): University of Texas Press (Original work published 1975).

Bearman, M., & Ajjawi, R. (2023). Can a rubric do more than grade? *British Journal of Educational Technology*, *54*, 5.

Boud, D., & Falchikov, N. (Eds.) (2007). *Rethinking assessment in higher education: Learning for the longer term*. London: Routledge.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43* (8), 1315-1325.

Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Westport (CT): Corwin Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34* (10), 906-911.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77* (1), 81-112.

Kasneci, E., Sesagiri Raamkumar, A., Kuchemann, S., Bannert, M., Anyadytiya, S., Zawacki-Richter, O., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *CHI Conference on Human Factors in Computing Systems*, 1-16.

Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *SSRN Preprint*.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, 100041.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31* (2), 199-218.

Panadero, E., & Jönsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129-144.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18* (2), 119-144.

Sansone, N. (in press). *The rAIse framework: Scaffolding metacognitive AI literacy through five-phase reflective practice*. British Journal of Educational Technology.

Vygotskij, L. S. (1978). *Mind in society: The development of higher psychological processes*. (M. Cole, V. John-Steiner, S. Scribner, E. Souberman, Eds.). Cambridge (MA): Harvard University Press.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41* (2), 64-70.

II.3
Dalla valutazione tradizionale alla valutazione simbiotica: analisi comparativa tra docenti e sistemi di IA nella valutazione accademica

Francesco Pio Sarcina
Università di Bari "Aldo Moro"[1], francesco.sarcina@uniba.it
Anna Maria Cuzzi
Università di Bari "Aldo Moro", anna.cuzzi@uniba.it
Michele Baldassarre
Università di Bari "Aldo Moro", michele.baldassarre@uniba.it

Abstract

Lo studio analizza 50 prove universitarie per confrontare valutazioni e feedback prodotti da un docente e da due modelli di IA (GPT e GEM), utilizzando rubriche condivise e analisi statistiche. I risultati mostrano un'elevata concordanza tra docente e ChatGPT, mentre Gemini adotta criteri più severi. La valutazione qualitativa evidenzia differenze nella chiarezza e nell'utilità dei feedback. La ricerca suggerisce che l'IA possa integrare, ma non sostituire, il giudizio umano nei processi valutativi.

Parole chiave: Intelligenza Artificiale; valutazione accademica; feedback personalizzato; prove semi-strutturate; valutazione formativa.

This study examines 50 university assessments to compare evaluations and feedback generated by a human instructor and two AI models (GPT and GEM), using shared rubrics and statistical analyses. Findings show strong alignment between the instructor and ChatGPT, while Gemini applies stricter grading. Qualitative data highlight differences in feedback clarity and usefulness. Results indicate that AI can complement, but not replace, human judgment, supporting more consistent and pedagogically meaningful assessment processes.

1 Sebbene gli autori abbiano condiviso la conduzione della ricerca *ivi* presentata e l'impostazione dell'articolo, si attribuisce a Francesco Pio Sarcina la scrittura dei paragrafi: 3, 3.1 e 3.2; ad Anna Maria Cuzzi la scrittura dei paragrafi: 2 e 4; a Michele Baldassarre la scrittura del paragrafo: 1.

Keywords: Artificial Intelligence; academic assessment; personalized feedback; semi-structured assessments; formative assessment.

1. Introduzione

Grazie alla capacità dei *Large Language Models* (LLM) di analizzare testi complessi e produrre giudizi coerenti, l'Intelligenza Artificiale (IA) sta trovando spazio nei processi di valutazione nell'istruzione superiore. Ricerche recenti evidenziano che gli strumenti generativi possono produrre valutazioni comparabili a quelle dei docenti in diversi compiti accademici, dalla correzione di saggi al giudizio su risposte aperte (Ishida et al., 2024; Lundgren, 2024). Un ulteriore aspetto significativo riguarda la capacità di questi modelli di produrre feedback personalizzati, che possono sostenere l'apprendimento autoregolato in contesti in cui la restituzione formativa del docente risulta spesso limitata o assente (Chiang et al., 2024; Matos et al., 2025).

Questi sviluppi sollevano tuttavia interrogativi sulla validità, sull'affidabilità e sulla trasparenza delle valutazioni fornite dalle IA, con particolare attenzione ai rischi di *bias* e interpretazioni distorte dei criteri valutativi (Stokkink, 2025; Wang et al., 2025). Alla luce di queste considerazioni, il presente studio mette a confronto modalità valutative effettuate da due modelli di IA generativa e da un docente, con l'obiettivo di verificare la coerenza del voto e la qualità del feedback offerto agli studenti. Tale confronto contribuisce a chiarire potenzialità e limiti dell'IA nei processi valutativi, offrendo elementi utili per una sua adozione pedagogicamente responsabile.

2. Quadro teorico

L'impiego dell'IA generativa nei processi valutativi universitari si colloca all'interno di un più ampio ripensamento dell'*assessment* in chiave digitale, che riguarda sia la misurazione degli apprendimenti, sia il ruolo del feedback nella progressione formativa. Le ricerche recenti evidenziano che i LLM, come ChatGPT e Gemini, sono capaci di analizzare testi accademici complessi e produrre valutazioni comparabili a quelle di esperti umani quando operano attraverso di rubriche analitiche o framework condivisi (Liang et al., 2025; Agostini et al., 2024; Potter et al., 2025). Ciò apre la possibilità di una automazione parziale delle pratiche valutative, migliorandone coerenza, trasparenza e tempesti-

vità. La generazione di feedback immediati, specifici e orientati al miglioramento, è una delle svolte tecnologiche che potrebbe consentire di superare una criticità strutturale tipica dell'istruzione superiore: l'impossibilità per molti docenti di fornire commenti personalizzati in tempi utili (Chiang et al., 2024; Létourneau et al., 2025). Tuttavia, la letteratura predica cautela, ricordando la necessità di presidiare la validità dei costrutti, l'allineamento con gli obiettivi formativi e il ruolo insostituibile del docente nel definire criteri, interpretare pedagogicamente le prove e controllare la qualità del giudizio (Elshall, 2025; Stokkink, 2025).

Trasparenza, tracciabilità, equità percepita e sostenibilità dei carichi di lavoro sono le condizioni imprescindibili per un'adozione responsabile dell'IA generativa nella valutazione universitaria (Kizilcec et al., 2024; Klimova et al., 2025; Matos et al., 2025), che potrà essere potenziata solo se pensata in ottica *human-centered* e inserita in solidi quadri metodologici ed etico-pedagogici.

3. Metodologia della ricerca

Questo studio ha utilizzato un approccio di ricerca quali-quantitativo. Sono state esaminate 50 prove scritte, selezionate casualmente da una popolazione di 160 elaborati prodotti da studenti del secondo anno del Corso di Laurea in Scienze della Formazione Primaria dell'Università di Bari. Gli studenti sono stati chiamati a rispondere a quattro quesiti di "Metodologia della ricerca educativa e didattica". Ogni elaborato è stato valutato tramite due rubriche (da 0 a 7 e da 0 a 8) e il punteggio finale assegnato è stato fino a 30/30, con l'eventuale possibilità di assegnare la lode (31/30).

Per quanto concerne l'analisi quantitativa, in una prima fase gli elaborati sono stati corretti in modo indipendente dal docente della disciplina, che ha assegnato i punteggi a ogni domanda e il voto finale. In un secondo momento, le stesse prove sono state valutate da due modelli di IA, precisamente tramite un GPT (versione personalizzata di ChatGPT) e un GEM (versione personalizzata di Gemini), che sono stati istruiti con le stesse rubriche utilizzate dal docente e i capitoli del manuale oggetto d'esame. I due strumenti di IA, addestrati con lo stesso *system prompt*, hanno restituito un punteggio e un feedback strutturato per domanda, oltre al voto finale e un feedback complessivo della prova. Infine, partendo da queste valutazioni, è stato effettuato uno studio statistico descrittivo attraverso analisi delle correlazioni, test di Friedman e test di Wilcoxon per confrontare le differenze tra i valutatori, seguiti da un'analisi Bland-Altman per esplorare la concordanza tra le valutazioni.

Per quanto concerne gli aspetti di natura qualitativa, ci si è concentrati su un duplice fronte: da un lato la valutazione della percezione dei feedback generati dai due modelli di IA da parte di un gruppo di cinque docenti-esperti, dall'altro la valutazione della percezione degli stessi da parte di un sotto-campione di studenti, che hanno potuto leggere i feedback alla loro prova. In entrambi i casi è stato somministrato un questionario con scala Likert con punteggi da 1 (totalmente in disaccordo) a 5 (pienamente d'accordo) per poter raccogliere le loro opinioni su criteri di chiarezza, coerenza, utilità, completezza, obiettività, applicabilità, equità e soddisfazione del giudizio.

3.1 Statistiche descrittive

È stato effettuato un confronto preliminare tra le medie (*Tab*. 1) dei punteggi assegnati dai tre valutatori (docente umano, GEM e GPT) per analizzare le tendenze generali nelle valutazioni e identificare eventuali differenze sistematiche (*Tab*. 2). Le medie dei punteggi sono state calcolate per ciascuna delle domande (Q1, Q2, Q3, Q4) e per il voto finale (somma dei punteggi delle quattro domande).

Valutatore	**Q1 (media)**	**Q2 (media)**	**Q3 (media)**	**Q4 (media)**	**Totale (media)**
Umano	6,5	5,4	6,8	7,1	26,75
Gemini	5,8	6,5	6,2	6,3	24,5
ChatGPT	6,6	5,7	6,9	7,2	27,76

Tab. 1 – Medie dei punteggi per ogni valutatore

Coppia di valutatori	**Q1**	**Q2**	**Q3**	**Q4**	**Totale**
Umano vs Gemini	0,7	-1,1	0,4	0,8	2,25
Umano vs ChatGPT	-0,1	-0,3	-0,1	-0,1	-1,01
Gemini vs ChatGPT	-0,8	-0,8	-0,7	-0,9	-3,26

Tab. 2 – Differenze tra le medie dei valutatori

I risultati evidenziano divergenze nelle medie dei punteggi assegnati dai tre valutatori. Si osservano due schieramenti: Docente Umano e ChatGPT nella

fascia alta, e Gemini più severo sul totale (24,5). ChatGPT risulta il più generoso (27,76), superando di circa un punto il docente (26,75; diff. +1,01). Le differenze Umano–ChatGPT per domanda restano contenute (Q1: -0,1; Q2: -0,3; Q3: -0,1; Q4: -0,1), suggerendo un giudizio spesso vicino ma leggermente più alto. La discrepanza maggiore emerge tra ChatGPT e Gemini (diff. complessiva=-3,26), con scarti sistematici in tutte le domande (da -0,7 a -0,9), soprattutto in Q1 dove Gemini scende a 5,8, distanziandosi dagli altri due. Va notato che in Q2 Gemini supera l'umano (6,5 rispetto a 5,4), invertendo trend. Al contrario, Q3 è l'area di minor divergenza tra valutatori. Questo indica che ChatGPT adotta un approccio valutativo mediamente più generoso rispetto al rigore di Gemini.

Il confronto delle mediane (*Tab*. 3) conferma un allineamento su Q2 (mediana=6) e segnala discrepanze più visibili su Q1 e Q4. In particolare, in Q4 la mediana di Gemini è 8,0 (massimo) ma la media è 6,3: un comportamento "tutto o niente" in cui molti 8 coesistono con punteggi molto bassi che trascinano la media. Nel complesso, queste evidenze descrittive delineano tendenze differenti e sollevano interrogativi su possibili divergenze sistematiche nei criteri valutativi.

Domanda	**Docente Umano**	**Gemini**	**ChatGPT**
Q1	6,0	5,0	6,0
Q2	6,0	6,0	6,0
Q3	7,5	7,0	7,0
Q4	7,75	8,0	7,0
Punteggio finale	26,75	24,5	27,25

Tab. 3 – Confronto delle mediane

Per approfondire queste differenze e analizzare la dispersione nelle valutazioni, si è passati all'analisi degli IQR (Interquartile Range). Il docente mostra la maggiore variabilità soprattutto in Q2 e nel punteggio finale (IQR=2,375 e 7,75), indicando un uso più esteso della scala. Le IA risultano invece più "piatte" sul totale (Gemini 4,75; ChatGPT 5,0), con voti più concentrati verso valori centrali. A livello di item, Gemini presenta l'IQR più ampio in Q1 (3,0) e valori elevati anche in Q3 (2,0), suggerendo una distribuzione più instabile; ChatGPT ha IQR più ridotti in Q1 (2,0) e Q2 (1,5), coerenti con una valutazione più uniforme. In Q4 la dispersione è identica per tutti (IQR=1,0). Questo passaggio

è stato essenziale per interpretare la coerenza interna dei valutatori e per comprendere come la variabilità, o la sua compressione, possa influire sulla qualità complessiva del processo valutativo.

Domanda	**Docente Umano (IQR)**	**Gemini (IQR)**	**ChatGPT (IQR)**
Q1	2,5	3,0	2,0
Q2	2,375	2,0	1,5
Q3	1,5	2,0	2,0
Q4	1,0	1,0	1,0
Punteggio finale	7,75	4,75	5,0

Tab. 4 – Confronto degli IQR

Il test di Friedman è stato eseguito per determinare se le differenze tra i valutatori fossero statisticamente significative. Questo test, che confronta i punteggi dei diversi valutatori su misure ripetute, ha mostrato che in Q1, Q2 e Q4 e nel punteggio finale le differenze tra i valutatori erano significative ($p<0,05$), mentre per la domanda Q3 ($p=0,2909$) non vi erano differenze statisticamente rilevanti (*Tab.* 5).

Domanda/Punteggio Finale	**p-*value***	**Significatività**
Q1	3.04e-06	Significativo
Q2	0,0267	Significativo
Q3	0,2909	Non significativo
Q4	0,00058	Significativo
Punteggio Finale	0,00181	Significativo

Tab. 5 – Test di Friedman per domanda e punteggio finale

Il test ha quindi confermato l'esistenza di divergenze significative tra i valutatori, rendendo necessaria una analisi *post-hoc* tramite il test di Wilcoxon (*Tab.* 6). Sul voto finale emerge convergenza tra Docente e ChatGPT ($p=0,8870$), pur con scarti su singole domande (Q2 $p=0,0157$; Q4 $p=0,0041$), concentrati nelle domande più critiche. Gemini, invece, si discosta: rispetto a ChatGPT la differenza sul totale è significativa ($p=0,0161$) e si concentra su Q1, Q2 e Q4;

rispetto al Docente la differenza complessiva è marginale (p=0,0971), nonostante la distanza su Q1 (p=0,0007). Il test ha dunque fornito una visione più dettagliata, indicando che Gemini tende a penalizzare più severamente gli studenti rispetto a ChatGPT, soprattutto nelle domande Q1, Q2 e Q4, mentre il docente e ChatGPT sono stati più allineati in molte delle risposte.

Domanda	**D. umano vs Gemini (p-*value*)**	**D. umano vs ChatGPT (p-*value*)**	**Gemini vs ChatGPT (p-*value*)**
Q1	0,0007	0,6195	1,22e-05
Q2	0,9014	0,0157	0,0024
Q3	0,2741	0,5888	0,4293
Q4	0,3615	0,0041	0,0011
Punteggio Finale	0,0971	0,8870	0,0161

Tab. 6 – Test di Wilcoxon per domanda e punteggio finale

Infine, l'analisi Bland-Altman è stata utilizzata per visualizzare la concordanza sul voto finale tra i valutatori. Nel confronto Docente Umano e Gemini (*Fig*. 1) le differenze risultano sistematiche e non casuali: bias Umano-Gemini=+1,87 e limiti di accordo molto ampi (–5,59; +9,32), con una dispersione prossima a 15 punti e concordanza eterogenea (Q1 e Q4 discrete; Q2 unica inversione; Q3 area più critica). Tra Docente Umano e ChatGPT (*Fig*. 2) l'allineamento è elevato: bias Umano-ChatGPT=–1,01, indicativo di lieve e costante generosità dell'IA, e limiti più contenuti (–7,86; +5,84), suggerendo accordo sul "chi" premiare. Al contrario, ChatGPT e Gemini (*Fig*. 3) mostrano la divergenza maggiore (quasi 3 punti): bias=+2,88 e LoA ampi (–3,34; +9,09), con ChatGPT più generoso e Gemini più severo, pur seguendo tendenze simili.

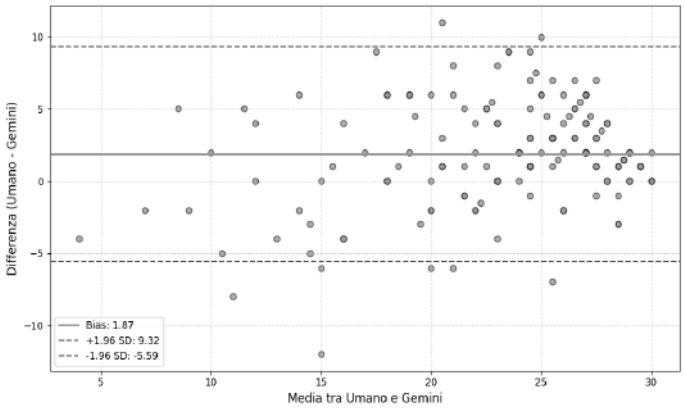


Fig. 1 – Limiti di concordanza D. umano vs Gemini sul punteggio totale

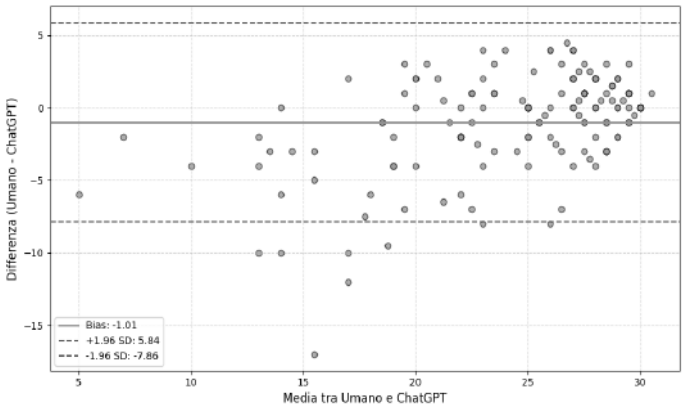


Fig. 2 – Limiti di concordanza D. umano vs ChatGPT sul punteggio totale

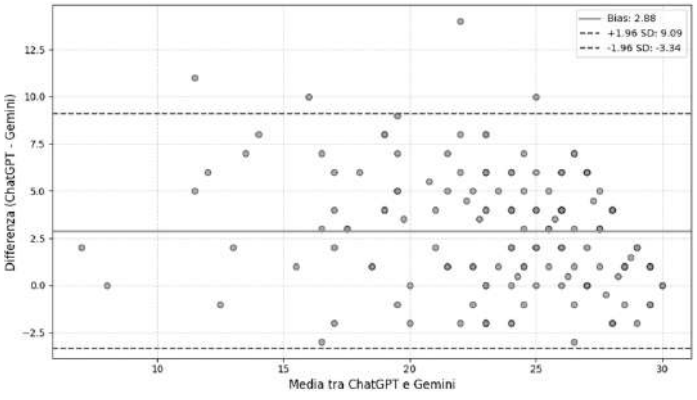


Fig. 3 – Limiti di concordanza ChatGPT vs Gemini sul punteggio totale

3.2 Analisi della qualità dei feedback dei modelli di IA

L'analisi dei risultati del questionario somministrato ai cinque docenti-esperti ha permesso di identificare chiaramente le differenze tra i due modelli di IA nell'ambito della valutazione dei feedback. In generale, i risultati del questionario mostrano che ChatGPT è stato valutato in modo più favorevole rispetto a Gemini, con punteggi più alti su molteplici dimensioni (*Tab*. 7).

Criteri	**ChatGPT (Media)**	**Gemini (Media)**
Chiarezza del feedback	3,6/5	4,0/5
Coerenza con i criteri di valutazione	4,0/5	4,0/5
Utilità nel migliorare l'apprendimento	3,8/5	4,0/5
Obiettività	3,6/5	3,8/5
Applicabilità dei feedback	4,0/5	4,2/5

Tab. 7 – Medie dei punteggi per criterio espresse dai docenti-esperti

Relativamente alla chiarezza, i docenti hanno trovato i feedback di Gemini (4,0) più comprensibili rispetto a quelli di ChatGPT (3,6). Sulla coerenza con i criteri di valutazione, entrambi i modelli hanno ottenuto lo stesso punteggio medio (4,0), dimostrando una consistente aderenza alle rubriche fornite. In termini di utilità, Gemini (4,0) ha superato lievemente ChatGPT (3,8), in quanto i suoi feedback sono percepiti come più efficaci nel supportare il miglioramento dell'apprendimento degli studenti; ChatGPT è risultato più orientato alla motivazione che alle strategie di studio. L'obiettività è stata giudicata alta per entrambi, sebbene Gemini (3,8) sia stato percepito come leggermente più severo e propenso a penalizzare risposte non pienamente conformi. Riguardo all'applicabilità, i feedback di Gemini (4,2) sono stati ritenuti più concreti e formativi rispetto a quelli di ChatGPT (4,0), percepiti come meno applicativi dagli esperti.

In conclusione, pur essendo simili per coerenza, Gemini è generalmente ritenuto più chiaro, utile e applicabile rispetto a ChatGPT, il cui focus primario sembra essere la motivazione. Entrambi sono obiettivi, ma Gemini è marginalmente più rigoroso.

L'analisi della qualità dei feedback forniti dai modelli di intelligenza artificiale ChatGPT e Gemini è stata estesa a un sotto-campione di dieci studenti, selezionati randomicamente tra quelli che hanno visto la loro prova valutata dai due sistemi di IA. I dati raccolti hanno permesso di esplorare i criteri di utilità nel

miglioramento, obiettività ed equità, chiarezza generale e rispetto agli errori commessi, soddisfazione rispetto ai feedback ricevuti, con risultati che hanno evidenziato diverse sfumature in alcuni aspetti indagati (*Tab*. 8).

Criteri	**ChatGPT (Media)**	**Gemini (Media)**
Utilità nel migliorare l'apprendimento	3,7/5	3,8/5
Spiegazione degli errori commessi	4,3/5	4,5/5
Obiettività ed equità	4,1/5	4,1/5
Chiarezza e comprensibilità	4,2/5	4,5/5
Soddisfazione rispetto al feedback	4,2/5	4,4/5

Tab. 8 – Medie dei punteggi per criterio espresse dagli studenti

I punteggi medi per l'utilità dei feedback mostrano una lieve preferenza per Gemini (3,8) rispetto a ChatGPT (3,7), sebbene due studenti abbiano assegnato punteggi molto bassi a entrambi i modelli.

Riguardo alla spiegazione e chiarimento degli errori, entrambi i modelli hanno ottenuto risultati eccellenti, con Gemini (4,5) quasi al massimo, confermando l'efficacia formativa di entrambe le IA nell'aiutare la comprensione degli sbagli. Equità e obiettività sono percepite allo stesso modo (4,1) per entrambi, ma la percezione della severità di Gemini risulta più eterogenea, pur riconoscendo l'oggettività generale dei feedback. La chiarezza è altamente apprezzata, specialmente quella di Gemini (4,5). I feedback di ChatGPT (4,2) sono comunque chiari, ma alcuni studenti li hanno trovati leggermente più tecnici. La soddisfazione generale è alta ma eterogenea. Gemini ha ottenuto un punteggio medio superiore (4,4) rispetto a ChatGPT (4,2), il quale ha ricevuto tre valutazioni molto basse. In base a quanto emerso, Gemini è generalmente preferito su quasi tutti gli aspetti, ma anche ChatGPT ottiene una percezione positiva. Per ottimizzare l'efficacia dei feedback, è cruciale considerare le differenze nelle esperienze individuali degli studenti.

4. Conclusioni e implicazioni pedagogiche

Sebbene ChatGPT e Gemini mostrino profili valutativi differenti, questo studio mira soprattutto a descrivere e quantificare, con strumenti statistici, l'ampiezza e la struttura degli scarti rispetto al docente, più che a proporre un modello pre-

scrittivo su *come* integrare l'IA nella valutazione, vista l'assenza di dati dedicati. I risultati evidenziano due traiettorie: ChatGPT mantiene un'elevata vicinanza al Docente Umano sul voto finale (26,75 vs 27,76; Wilcoxon p=0,887; bias −1,01), mentre Gemini risulta più severo e più variabile, con disaccordo strutturale (bias Umano–Gemini +1,87) e una distanza significativa rispetto a ChatGPT (p=0,0161; bias +2,88). Le differenze sono concentrate in Q1, Q2 e Q4, mentre Q3 non risulta significativa nel test di Friedman, configurandosi come area di massimo accordo; coerentemente, i limiti di accordo risultano più ampi nei confronti con Gemini (Bland-Altman). L'analisi degli IQR mostra inoltre che il docente utilizza maggiormente la scala sul totale (IQR 7,75), mentre le IA tendono a comprimere la variabilità (4,75–5,0), suggerendo una maggiore "piattezza" del giudizio.

Sul piano formativo, i questionari mostrano che i modelli non sono intercambiabili: Gemini è percepito mediamente più chiaro e applicabile da docenti-esperti e studenti, mentre ChatGPT tende a produrre feedback più discorsivi e di supporto, talvolta meno operativi e con soddisfazione più eterogenea.

Alla luce di questi risultati, parlare di IA come strumento complementare per la valutazione tradizionale va inteso in senso sostenibile: non come tripla correzione, ma come uso dell'IA per una lettura controllata o la generazione di feedback preliminari, lasciando al docente la decisione finale, la responsabilità valutativa e il controllo di coerenza con rubriche e obiettivi formativi. In pratica l'IA può essere impiegata per segnalare *outlier* e supportare la revisione umana. Resta necessario testare empiricamente, in studi futuri, *workflow* ibridi (rubriche, *prompt* e controlli di qualità) e valutarne effetti su tempi di correzione, percezione di equità e apprendimenti, così da collegare il confronto tra modelli a esiti educativi misurabili.

L'obiettivo potrebbe essere quello di approdare verso una "valutazione simbiotica", intesa come il processo integrato in cui l'IA e il giudizio umano collaborano in modo circolare, unendo la capacità computazionale e analitica della macchina alla sensibilità etica, al pensiero critico e alla comprensione del contesto tipiche dell'uomo per generare un giudizio più accurato, equo e profondo di quanto farebbero se valutassero separatamente.

Riferimenti bibliografici

Agostini, D., Romano, M., & Bertoldi, N. (2024). *Are Large Language Models Capable of Assessing Educational Content? Educational Technology Research and Development.* https://doi.org/10.6093/2284-0184/10671.

Chiang, C.-H., Chen, W.-C., Kuan, C.-Y., Yang, C., & Lee, H.-Y. (2024). *Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2489-2513. https://doi.org/10.18653/v1/2024.emnlp-main.146.

Elshall, A. S. (2025). *Balancing AI-assisted learning and traditional assessment: the FACT assessment in environmental data science education. Frontiers in Education*, 10. https://doi.org/10.3389/feduc.2025.1596462.

Ishida, T., Liu, T., Wang, H., & Cheung, W. K. (2024). *Large Language Models as Partners in Student Essay Evaluation.* arXiv preprint. https://doi.org/10.48550/arXiv.2405.18632.

Kizilcec, R. F., Holstein, K., & Aleven, V. (2024). *Perceived impact of generative AI on assessments: International survey of educators and students. Computers and Education: Artificial Intelligence*, 7. https://doi.org/10.1016/j.caeai.2024.100269.

Klimova, B., Pikhart, M., & Polakova, P. (2025). *Exploring the effects of artificial intelligence on student and teacher well-being in higher education: A mini-review. Frontiers in Psychology*, 11830699. https://doi.org/10.3389/fpsyg.2025.1498132.

Létourneau, A., Bouchet, F., & Lavoué, E. (2025). *A systematic review of AI-driven intelligent tutoring systems in K-12 education: Effects on learning and performance. BMC Medical Education*, 12078640. https://doi.org/10.1186/s12909-025-05678-1.

Liang, J., Stephens, J. M., & Brown, G. T. L. (2025). *A systematic review of the early impact of artificial intelligence on higher education curriculum, instruction, and assessment. Frontiers in Education*, 10. https://doi.org/10.3389/feduc.2025.1522841.

Lundgren, M. (2024). *Large Language Models in Student Assessment: Comparing ChatGPT and Human Graders.*arXiv preprint. https://doi.org/10.48550/arXiv.2406.16510.

Matos, T., Santos, A., & Silva, R. (2025). *A systematic review of artificial intelligence applications in educational assessment: Focus on effectiveness and challenges. Computers and Education: Artificial Intelligence.*https://doi.org/10.1016/j.dajour.2025.100571.

Potter, A., Liu, C., & Zhang, Y. (2025). *Assessing academic language in tenth grade essays using natural language processing tools. System*, 129. https://doi.org/10.1016/j.system.2025.103278.

Stokkink, P. (2025). *The Impact of AI on Educational Assessment: A Framework for Constructive Alignment.* arXiv preprint. https://doi.org/10.48550/arXiv.2506.23815.

Wang, J., Li, Y., Chen, X., & Zhang, L. (2025). *The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. Humanities and Social Sciences Communications*, 12. https://doi.org/10.1038/s41599-025-04787-y.

II.4
Come gli studenti universitari valutano la risoluzione di un task generata da ChatGPT

Giovannina Albano
Università degli Studi di Salerno, via Giovanni Paolo II, 132, 84084, Fisciano (SA), Italy
galbano@unisa.it
Anna Pierri
Università Giustino Fortunato, 82100 Benevento, Italy
a.pierri@unifortunato.eu
Agnese Ilaria Telloni
Università degli Studi di Macerata. 62100 Macerata, Italy
agnese.telloni@unimc.it

Abstract

Lo studio si basa sul design di un esperimento mirato a investigare il potenziale ruolo dell'intelligenza artificiale come strumento educativo per sviluppare il pensiero critico e le competenze argomentative degli studenti. In particolare, vengono analizzate le valutazioni fatte da studenti universitari sulla risoluzione di un task di Algebra Lineare generata da ChatGPT. Sullo sfondo di un quadro teorico che integra i costrutti della teoria antropologica della didattica e le massime conversazionali di Grice, si conduce un'analisi a priori della risoluzione del quesito generata da ChatGPT e si confronta tale analisi con le valutazioni fatte da 54 matricole di Ingegneria. I risultati suggeriscono che le valutazioni degli studenti si discostano dall'analisi a priori soprattutto per quanto concerne i collegamenti logici e la dimensione comunicativa.

Parole chiave: intelligenza artificiale generativa; studenti universitari; valutazione di processi risolutivi; metacognizione; argomentazione.

The study is based on the design and implementation of a teaching experiment aimed at investigating the potential role of artificial intelligence as an educational tool for developing students' critical thinking and argumentation skills. In particular, it analyzes university students' evaluations of the solution to a Linear Algebra task generated by ChatGPT. Within a theoretical framework integrating the Anthropological Theory of Didactics and Grice's conversational maxims, an a priori analysis of the solution produced by ChatGPT is conducted and com-

pared with the evaluations made by 54 Engineering freshmen. The results suggest that students' evaluations diverge from the a priori analysis especially with regard to logical connections and the communicative dimension.

Keywords: artificial intelligence; university students; evaluation of solving processes; metacognition; argumentation.

1. Introduzione

L'integrazione dell'intelligenza artificiale generativa (GenAI) sta rapidamente trasformando l'istruzione, incidendo su come studenti e docenti interagiscono con contenuti e risorse (Richard, Pilar V lez & Van Vaerenbergh, 2022).

L'ampio uso di ChatGPT interroga il mondo della ricerca didattica: l'uso di quadri pedagogici teorici è ancora limitato e le sfide e implicazioni didattiche sono ancora troppo poco investigate (Zawacki-Richter, Marín, Bond et al., 2019; Pepin, Buchholtz & Salinas-Hernández, 2025). Dal punto di vista didattico, è necessario equilibrare l'uso della GenAI con lo sviluppo del pensiero autonomo. È importante che i docenti comprendano la *ratio* alla base delle risposte di ChatGPT, per poterne sfruttare le potenzialità e allo stesso tempo governare i limiti.

In questa direzione, il presente studio parte dall'analisi a priori del comportamento di ChatGPT nel rispondere a una domanda con una particolare struttura logica (Albano, Pierri & Telloni, 2025), e investiga come gli studenti valutano le risposte di ChatGPT rispetto alle criticità emerse nell'analisi a priori.

2. Quadro teorico di riferimento

Albano, Swidan e Pierri (2023) hanno sviluppato un modello per l'analisi di risposte argomentative a task matematici, guardando alle produzioni degli studenti come segmenti di conversazioni tra lo studente, che produce il testo scritto, e il docente che deve leggere tale testo. Per questo motivo il modello sviluppato considera sia le componenti matematiche sia quelle comunicative, integrando ed estendendo la dimensione prasseologica sviluppata all'interno della Teoria Antropologica della Didattica (ATD, Bosch & Gascón, 2014) con la dimensione pragmatica di Grice (1975).

Sul piano matematico, la dimensione ATD interpreta la risoluzione di un compito come una prasseologia, intesa come l'insieme interrelato di prassi (la

componente operativa) e logos (la componente razionale e giustificativa). In questa prospettiva, l'analisi mira a identificare, all'interno di ciascuna risposta, gli elementi di tecnica, tecnologia e teoria, nonché la catena logico-deduttiva che collega i passaggi argomentativi. La solidità di questa catena funge da indicatore della capacità del rispondente di coordinare diversi registri semiotici e di mantenere coerenza tra le procedure operative e la loro giustificazione concettuale. Sul piano comunicativo, la dimensione griceana valuta il rispetto delle quattro massime conversazionali – quantità, qualità, relazione e maniera – che permettono di analizzare coerenza, chiarezza e pertinenza del discorso matematico prodotto. Ogni risposta viene quindi esaminata in termini di completezza delle informazioni, correttezza dei contenuti, pertinenza rispetto alla domanda e chiarezza espositiva. Sulla base di quanto detto, il modello sviluppato da Albano, Swidan e Pierri si sintetizza nella tabella mostrata nella Figura 1.

	Componenti dell'ATD ampliato			
Massime di Grice	Tecnica	Teoria	Tecnologia	Catena logica
Quantità (abbastanza/ridondante/scarno)				
Qualità (vero/supportato da evidenze/falso)				
Relazione (pertinente/non pertinente)				
Maniera (chiaro/ambiguo/non chiaro)				

Fig. 1 – Modello di analisi di risposte argomentative a task matematici

3. Metodo

3.1 Il contesto

Lo studio è stato condotto con due gruppi di matricole di Ingegneria, iscritte a un corso di Algebra Lineare nel primo semestre dell'anno accademico 2025-26. Un gruppo è composto da 34 studenti di Ingegneria Gestionale dell'Università Politecnica delle Marche; l'altro gruppo comprende 20 studenti di Ingegneria Edile-Architettura dell'Università di Salerno. La partecipazione allo studio è avvenuta esclusivamente su base volontaria. I contenuti principali trattati nei due corsi sono gli stessi, così come le tipologie di quesiti a cui gli studenti sono stati esposti (domande aperte, chiuse, argomentazioni sulla correttezza di item proposti come risposta) durante il semestre.

3.2 La progettazione dell'esperimento

Il design dell'esperimento prevede le seguenti fasi di attività:

1. scelta di un task e sottomissione a ChatGPT;
2. analisi a priori della risposta fornita da ChatGPT;
3. sottomissione agli studenti della risposta di ChatGPT con richiesta di valutazione secondo alcuni criteri definiti;
4. analisi delle valutazioni degli studenti.

I task sottoposti a ChatGPT sono domande aperte su argomenti di Algebra Lineare, estratte da quelle utilizzate nella valutazione formativa o sommativa nei nostri corsi per matricole di ingegneria. La struttura delle domande è la seguente:

1. *Domanda principale*: consiste nel seguente testo "Dato il seguente quesito, giustifica in modo chiaro e completo la correttezza o meno dell'item proposto come possibile risposta";
2. *Quesito*: consiste in una domanda aperta;
3. *Item proposto come risposta*: consiste in una possibile risposta sintetica al quesito.

Allo studente non si chiede di rispondere al quesito, ma di valutare la correttezza dell'item proposto come risposta, fornendo un'argomentazione chiara e completa. La Figura 2 mostra il task oggetto dello studio presentato in questo paper.

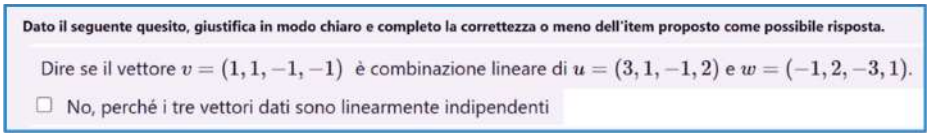


Fig. 2 – Task sottoposto a ChatGPT

Osserviamo che per risolvere il task basta verificare la lineare indipendenza dei tre vettori dati, attraverso il rango della matrice che li ha come righe o colonne (che risulta pari a 3). Ciò infatti consente di concludere la verità dell'*Item proposto come risposta*, dal momento che la lineare indipendenza dei tre vettori equivale a dire che nessuno dei tre vettori è una combinazione lineare degli altri due.

3.3 Analisi a priori della risposta di ChatGPT al task sottomesso

Di seguito presentiamo l'analisi della risposta di ChatGPT utilizzando il modello di Albano, Swidan e Pierri descritto in precedenza. Nel rispondere al task proposto, ChatGPT parte col risolvere il *Quesito*, ignorando l'*Item proposto come risposta*. Dalla Figura 3, si può notare che l'approccio evolve secondo i seguenti passi: riformulazione della domanda principale tramite l'uso implicito della definizione di combinazione lineare; passaggio dall'equazione vettoriale alle equazioni scalari, con l'aggiunta di un riferimento esplicito alle componenti dei vettori coinvolti; risoluzione del sistema per sostituzione; conclusione che gli scalari desiderati non esistono. Val la pena di notare che non solo non è necessaria la risoluzione del sistema, bastando solo la compatibilità, ma la tecnica utilizzata (sostituzione) non è adeguata al livello universitario.

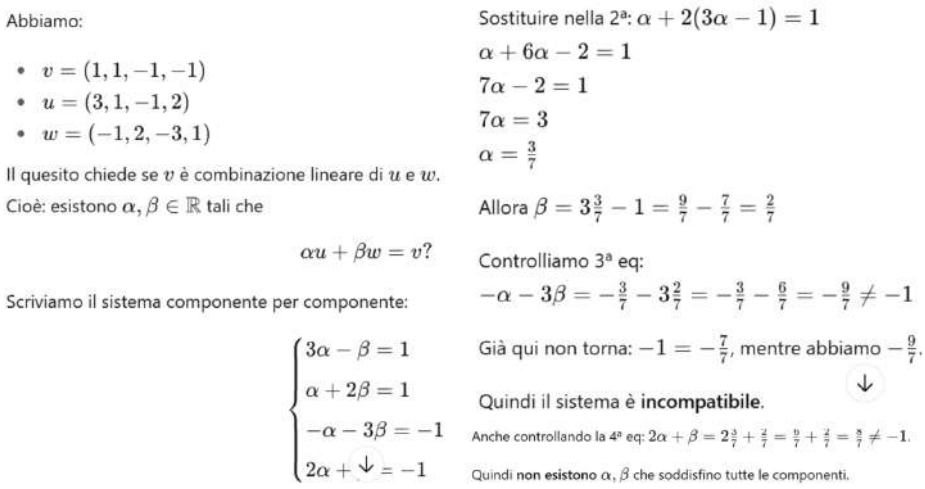


Fig. 3 – Risoluzione del task da parte di ChatGPT - parte 1

Solo nella parte finale della risoluzione (Figura 4), ChatGPT prende in considerazione l'*Item proposto come risposta*. Qui possiamo osservare un chiaro errore logico. Da un lato, afferma correttamente che l'indipendenza lineare dei tre vettori significa che nessuno di essi è una combinazione lineare degli altri. Dall'altro, non riesce a riconoscere l'equivalenza tra "nessuno dei tre è" e "ciascuno dei tre non è" (una combinazione lineare degli altri due). Ancora più sorprendentemente, non viene riconosciuto che stabilire che ciascuno dei tre non è una combinazione lineare degli altri due implica necessariamente, in particolare, che *v*

non è una combinazione lineare di u e w. Infine, osserviamo che non viene presa una posizione sulla verità o falsità dell'affermazione secondo cui i tre vettori dati sono linearmente indipendenti.

Conclusione:

Il vettore v **non è combinazione lineare** di u e w.
Ma il motivo non è "perché i tre vettori sono linearmente indipendenti" — quello è un concetto diverso.
Dire che tre vettori in $\mathbb{R}^4$ sono linearmente indipendenti significa che **nessuno dei tre** è combinazione lineare degli altri due. Ma qui stiamo chiedendo solo se v è combinazione di u e w.

La risposta corretta è:

- **No, perché non esistono scalari α, β tali che $\alpha u + \beta w = v$.**

Fig. 4 – Risoluzione del task da parte di ChatGPT - parte 2

L'analisi proposta può essere sintetizzata nella seguente tabella (Figura 5).

	Componenti dell'ATD ampliato			
Massime di Grice	**Tecnica**	**Teoria**	**Tecnologia**	**Catena logica**
Quantità (abbastanza/ridondante/scarno)	Abbastanza	Scarno	Scarno	Scarno
Qualità (vero/supportato da evidenze/falso)	Vero	Falso	Supportato da evidenze	Falso
Relazione (pertinente/non pertinente)	Pertinente	Pertinente	Pertinente	Pertinente
Maniera (chiaro/ambiguo/non chiaro)	Non chiaro	Non chiaro	Non chiaro	Non chiaro

Fig. 5 – Analisi della risoluzione del task generata da ChatGPT

3.4 Il task sottomesso agli studenti

Agli studenti è stato chiesto di valutare la risoluzione del task svolta da ChatGPT, concentrandosi sulle procedure utilizzate, sulla loro giustificazione teorica e sui collegamenti logici fra le varie parti. Per ciascuno di questi aspetti, sono state proposte 4 domande a risposta chiusa (scelta multipla, Figura 6). Ogni domanda corrisponde a una colonna della matrice in Figura 1, cioè a una dimensione del modello ATD ampliato. Per facilitare gli studenti, le dimensioni "Tecnologia" e

"Teoria" sono state unite nella richiesta relativa alle "giustificazioni teoriche". Le opzioni fornite come risposte riproducono le possibili scelte sulle massime di Grice. Infine, agli studenti è stato chiesto di giustificare le valutazioni date in una risposta aperta, in modo che potessero concentrarsi sugli aspetti ritenuti più rilevanti (Connelly & Clandinin, 1990; Bell, 2002).

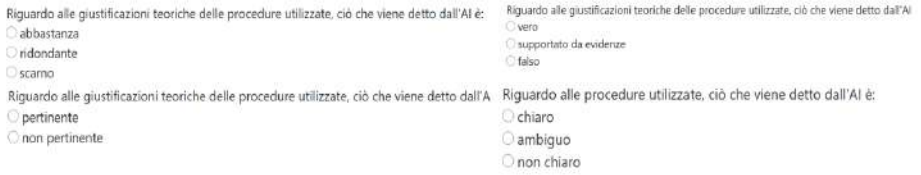


Fig. 6 – Le 4 domande a scelta multipla relative alle "procedure"

3.5 Raccolta dei dati e strumenti di analisi

Poiché i task sono stati sottomessi attraverso la risorsa "Sondaggio" di Moodle, contenente sia domande chiuse (scelta multipla) sia domande aperte, i dati sono stati raccolti attraverso l'archiviazione delle risposte date dagli studenti e i relativi report in piattaforme.

Abbiamo analizzato i dati sia da un punto di vista quantitativo che qualitativo.

Dal punto di vista quantitativo, abbiamo focalizzato l'attenzione sui tre elementi: procedure, giustificazione e collegamenti logici, riportando le percentuali di risposte in Figura 6 per i due campioni.

Dal punto di vista qualitativo, lo studio si fonda sull'assunto che i ricercatori ricavino significato e comprensione interpretando i fenomeni attraverso le parole dei narratori (Bell, 2002). Pertanto, abbiamo analizzato le risposte alla domanda aperta nel quadro delle massime di Grice. Abbiamo posto particolare attenzione alla selezione di alcuni esempi paradigmatici giustificativi dei dati quantitativi ottenuti.

4. Risultati e discussione

L'analisi delle risposte fornite dai due campioni evidenzia alcune tendenze significative nel modo in cui gli studenti valutano la risoluzione del task generata da ChatGPT (Tabella 1). In particolare, emerge una soddisfazione diffusa ri-

spetto alle procedure adottate e all'articolazione dei collegamenti logici, mentre risultano più critiche le valutazioni relative alle giustificazioni teoriche.

Per quanto riguarda le procedure, la maggioranza degli studenti le giudica chiare e comprensibili. Anche quando vengono segnalate ambiguità, queste non sembrano compromettere in modo decisivo la valutazione complessiva: molti studenti riconoscono comunque una familiarità nei passaggi operativi.

In merito alle giustificazioni teoriche, se una parte consistente degli studenti le giudica "vere" o "supportate da evidenze", alcuni ne rilevano invece i limiti, in particolare in termini di quantità.

Infine, l'analisi dei collegamenti logici mostra una percezione generalmente positiva della risoluzione di ChatGPT, nonostante siano state rilevate alcune criticità in termini di coerenza: alcuni studenti colgono l'incompletezza della risoluzione di ChatGPT o il fatto che non risponda alla domanda principale.

	Procedure		**Giustificazioni teoriche**		**Collegamenti logici**	
	I camp.	II camp.	I camp.	II camp.	I camp.	II camp.
Scarno	18%	10%	47%	55%	21%	35%
Abbastanza	76%	80%	47%	3%	18%	10%
Ridondante	6%	10%	6%	15%	56%	55%
Falso	0%	5%	24%	25%	12%	15%
Supportato da evidenze	47%	30%	26%	4%	35%	25%
Vero	47%	65%	50%	35%	47%	60%
Non pertinente	3%	5%	26%	35%	24%	30%
Pertinente	94%	95%	74%	65%	71%	70%
Oscuro	0%	0%	9%	10%	0%	0%
Ambiguo	35%	45%	44%	60%	44%	50%
Chiaro	59%	55%	41%	30%	50%	50%

Tab. 1 – Le valutazioni degli studenti attraverso le domande a scelta multipla

I due campioni hanno fornito una valutazione piuttosto omogenea della risoluzione generata da ChatGPT. In particolare, le procedure e i collegamenti logici sono stati valutati complessivamente in modo positivo. Infatti, la mag-

gioranza degli studenti ha giudicato **abbastanza** (76% del primo campione e 80% del secondo), **vero** (47% del primo campione e 65% del secondo) **e pertinente** (94% del primo campione e 95% del secondo) quanto detto dall'AI sulle procedure. Gli stessi giudizi sono emersi per i collegamenti logici, che sono stati ritenuti **abbastanza** dal 56% del primo campione e dal 55% del secondo, **veri** dal 47% del primo campione e dal 60% del secondo e **pertinenti** dal 71% del primo campione e dal 70% del secondo. Anche la comunicazione delle procedure e dei collegamenti logici è stata ritenuta soddisfacente: almeno il 50% di ciascuno dei campioni ha giudicato **chiaro** quanto detto dall'AI. Alcuni studenti (35% del primo campione e 45% del secondo) hanno invece classificato come **ambigue** le procedure, segnalando che talvolta i passaggi non risultano coerenti con la richiesta del task. Per quanto concerne i collegamenti logici nella risoluzione di ChatGPT, emerge che il 24% del primo campione e il 30% del secondo li ha ritenuti **non pertinenti**; inoltre, il 12% del primo campione e il 15% del secondo li ha dichiarati **falsi**. Si evidenzia dunque che la principale fragilità della risposta dell'AI non risiede nella tecnica o nelle definizioni utilizzate, ma nella capacità di costruire una catena argomentativa solida e coerente.

Generalmente più critiche sono state le valutazioni da parte degli studenti delle giustificazioni teoriche, dichiarate **scarne** (47% del primo campione e 55% del secondo), **false** (circa il 25% in ciascun campione) e **ambigue** (44% del primo campione e 60% del secondo). Una parte degli studenti ha rilevato, ad esempio, che l'AI non articola correttamente il legame logico tra l'indipendenza lineare fra vettori e l'impossibilità di esprimere uno di essi come combinazione lineare degli altri.

L'analisi condotta ha fatto emergere alcuni aspetti (in grassetto in Tabella 1) su cui la valutazione della risoluzione generata da ChatGPT fatta dagli studenti si discosta da quella in Figura 5.

	Procedura	Giustificazioni teoriche	Collegamenti logici
Quantità	Abbastanza	Scarno	**Abbastanza**
Qualità	Vero	**Vero**	**Vero**
Relazione	Pertinente	pertinente	Pertinente
Maniera	**Chiaro**	ambiguo	**Chiaro**

Tab. 2 – Matrice che risulta dalle valutazioni degli studenti dei due campioni

In relazione a tali aspetti, sono stati realizzati i grafici nelle Figure 7, 8 e 9.

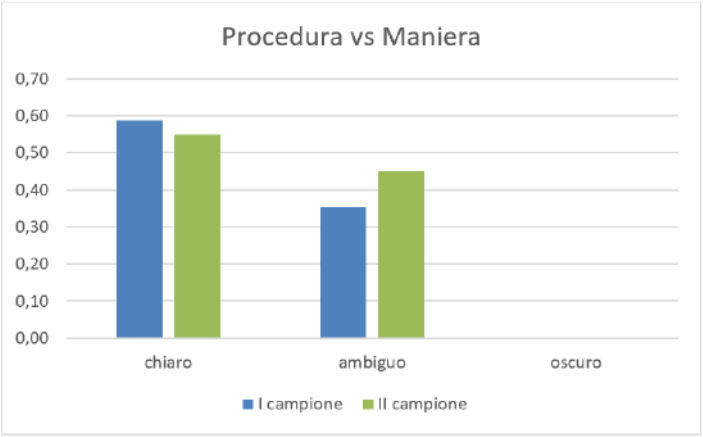


Fig. 7 – La valutazione delle procedure secondo la Maniera

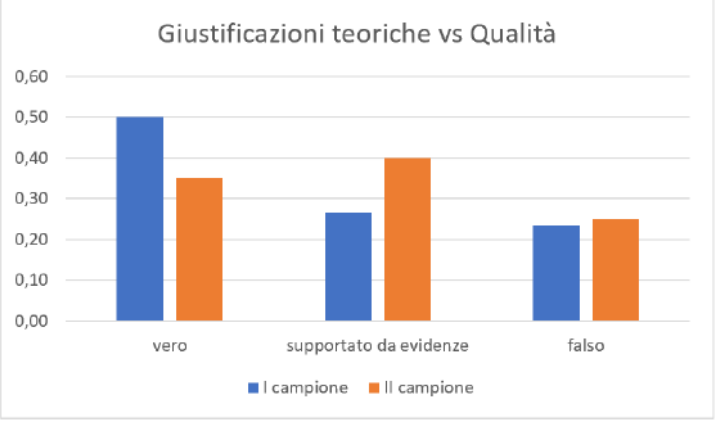


Fig. 8 – La valutazione delle giustificazioni teoriche secondo la Qualità

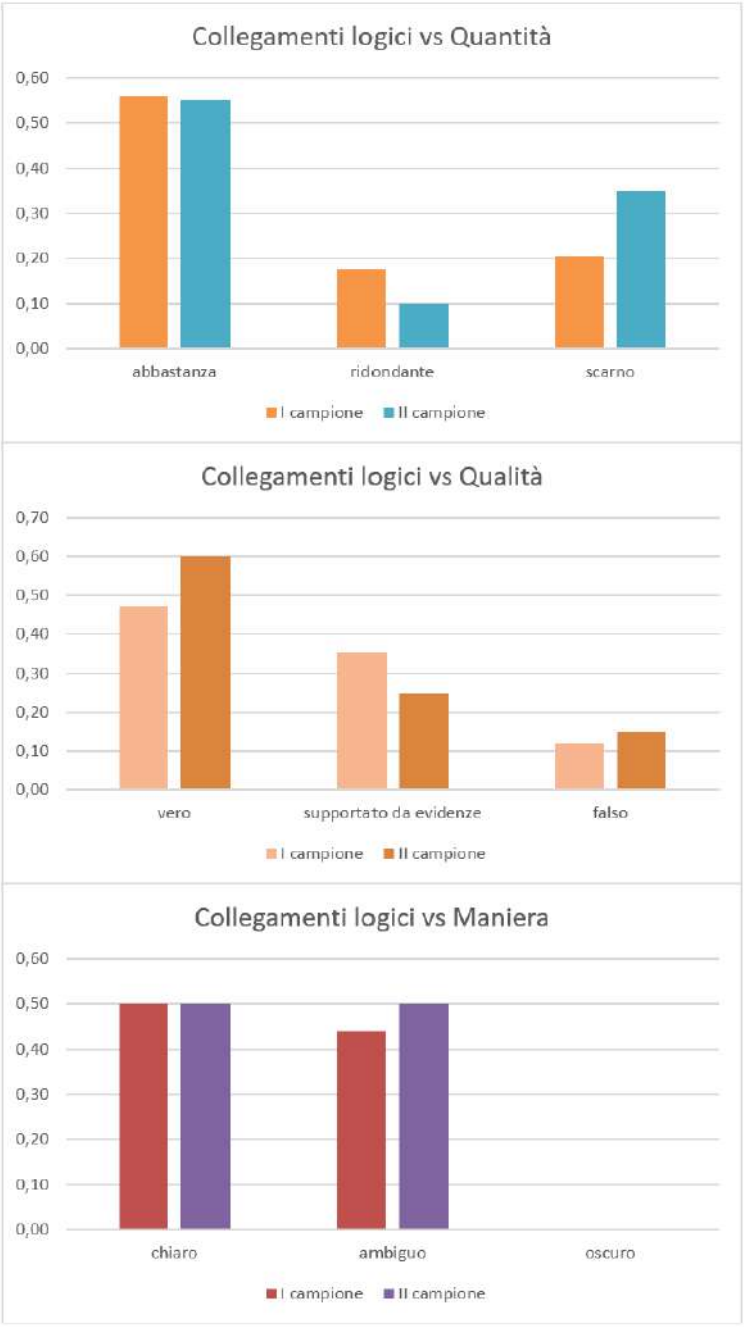


Fig. 9 – La valutazione dei collegamenti logici

In particolare, le procedure sono state ritenute **chiare**, in quanto "*sostenute da passaggi algebrici"* (S32), che *"sono eseguiti in modo lineare e comprensibile"* (S5). Alcune risposte sottolineano che *"la procedura è chiara ma non pertinente alla domanda principale"* (S21). Questo mostra un punto chiave: gli studenti distinguono con difficoltà tra "procedura eseguita correttamente" e "procedura adeguata alla struttura logica del task".

Differentemente dalla nostra analisi a priori, le giustificazioni teoriche sono ritenute **vere**, "*pertinenti e supportate da calcoli e definizioni corrette*" (S34). Tuttavia, alcuni studenti hanno rilevato elementi critici, in linea con la nostra valutazione, dichiarando che nella risoluzione di ChatGPT "*manca profondità teorica*" (S1), o che i riferimenti sono talvolta *"insufficienti e confusi"* (S13), ad esempio perché l'AI *"non esplicita l'equazione vettoriale iniziale"* (S12). Diversi studenti hanno colto la criticità nella risoluzione di Chat relativamente al legame concettuale tra indipendenza lineare e impossibilità di esprimere un vettore come combinazione lineare degli altri due. Ad esempio, S8 ha rilevato che "*l'affermazione chiave della critica* (*"quello è un concetto diverso",* cfr. Figura 4) *è teoricamente imprecisa; il fatto che v non sia combinazione lineare di u e w è esattamente la condizione che porta all'indipendenza lineare dell'insieme, la relazione è quindi di conseguenza logica"*. Questa inesattezza è segnalata spesso anche nella valutazione della catena logica (S30: "*ritengo che i collegamenti logici non siano ben definiti e risultano essere un po' ambigui dato che entra in campo sempre il concetto di lineare indipendenza non ben connesso con il procedimento e la giusta risposta al quesito*"). Proprio i collegamenti logici rappresentano la dimensione in cui la nostra analisi a priori e la valutazione degli studenti mostrano le differenze più marcate (Tabella 2, Figura 9). S23 ha dichiarato che "*i collegamenti sono chiari e pertinenti"*, S2 ha osservato che *"ogni passaggio porta naturalmente al successivo. Io credo che [i collegamenti logici] siano pertinenti e basati sui risultati ottenuti e li trovo chiari da seguire"*. Solo pochi studenti hanno colto le criticità logiche nella risoluzione di ChatGPT, fra cui il fatto che prima il chatbot risolve il quesito, poi valuta l'item, senza collegare le due parti (S21: "*la motivazione iniziale dell'esercizio non è pertinente con lo svolgimento dell'esercizio, perché il procedimento risolve un'altra cosa*"; S15: *"la risposta fornita non soddisfa la richiesta della domanda, è vaga in ciò che risulta essere il concetto principale"*).

Inoltre, l'AI gestisce in modo problematico la quantificazione (cfr. sezione 3.4), ma nessuno studente ha segnalato questa criticità.

L'analisi svolta ha fatto emergere alcuni giudizi entusiastici sullo svolgimento generato da ChatGPT, in particolare riguardo alla dimensione comunicativa. S4 ha dichiarato che: "*È tutto giusto e corretto riguardo le procedure utilizzate e i collegamenti logici tra le parti della risoluzione. Anche per quanto riguarda la giu-*

stificazione teorica l'AI è stato impeccabile!"; S26 ha osservato che "*La risoluzione è corretta, chiara e concettualmente rigorosa. Le procedure di calcolo sono eseguite con precisione".*

5. Conclusione

Lo studio ha investigato le caratteristiche della risoluzione generata da ChatGPT di un quesito di Algebra Lineare incentrato sull'implicazione logica e ha confrontato l'esito dell'analisi a priori di tale risoluzione con quella che ne hanno fatto 54 matricole di Ingegneria di due diversi atenei. È emerso che la discrepanza più significativa fra l'analisi a priori e le valutazioni degli studenti riguarda i collegamenti logici e la dimensione comunicativa. In particolare, rispetto ai primi, gli studenti tendono a non cogliere le fallacie presenti nella risoluzione generata da ChatGPT e messe in luce nell'analisi a priori. Rispetto alla seconda, le valutazioni degli studenti sono generalmente più positive rispetto all'analisi a priori. Talvolta addirittura l'apprezzamento della risoluzione fornita da ChatGPT appare entusiastico. L'apprezzamento del processo risolutivo generato dall'AI e in particolare della modalità di presentare gli argomenti merita un approfondimento. Esso potrebbe essere dovuto al fatto che alcuni concetti sono semplificati, la struttura della risposta è articolata in passaggi facilmente riconoscibili, il lessico è più vicino al linguaggio a cui gli studenti sono abituati rispetto a quello della matematica ufficiale. Inoltre, l'uso di ChatGPT come tutor o interlocutore abituale da parte degli studenti, sia per argomenti di studio, sia per altri argomenti, potrebbe avere un'incidenza sulla loro valutazione. In futuro vorremmo approfondire l'indagine su questo aspetto, cercando di comprendere in che modo l'uso dell'AI influenzi e possa influenzare la competenza comunicativa in relazione ai contenuti matematici.

Un ulteriore sviluppo dello studio qui presentato sarà l'uso del design dell'esperimento presentato nella sezione 4.2 come modello metodologico per coinvolgere gli studenti in attività di meta-valutazione, volte a sostenere lo sviluppo del senso critico e della competenza argomentativa. Uno degli aspetti più rilevanti emersi nell'ambito delle ricerche che stiamo conducendo (Albano et al., 2025), infatti, è che alcune caratteristiche delle risoluzioni generate dall'intelligenza artificiale, quali la tendenza a privilegiare approcci procedurali e il non determinismo (cioè il fatto che a uno stesso input non corrisponde sempre il medesimo output), pur essendo delle criticità sotto certi punti di vista, rappresentano invece un'opportunità molto significativa per favorire lo sviluppo di capacità critiche e di consapevolezza sui significati matematici da parte degli studenti.

Riferimenti bibliografici

Albano, G., Swidan, O., & Pierri, A. (2023). A model for analyzing the explanatory writing of undergraduate students when solving mathematical tasks. *International journal of mathematical education in science and technology*, *54*(2), 180-194.

Albano, G., Pierri, A., & Telloni, A.I. (2025). ChatGPT in University Education: how does the tool perform on mathematical questions with a specific structure? Proceedings of HELMeTO 2025, Naples September 23-25, 2025 (to appear).

Bell, S. (2002). Narrative inquiry: More than just telling stories. *TESOL Quarterly*, 36(2), 207-213.

Bosch, M., & Gascón, J. (2014). Introduction to the Anthropological Theory of the Didactic (ATD). In A. Bikner-Ahsbahs, & S. Prediger (Eds.), *Networking of theories as a research practice in Mathematics education* (pp. 67-83). Advances in Mathematics Education. Springer.

Connelly, F. M., & Clandinin, D.J. (1990). Stories of experience and narrative inquiry. Educational Researcher, 19 (5), 2-14.

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. Morgan (Eds.), *Syntax and semantics*, Vol. 3 (pp. 41-58). Academic Press

Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025) A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education*, 11, 9-41.

Richard, P. R., Pilar V lez, M., & Van Vaerenbergh, S. (2022). *Mathematics education in the age of artificial intelligence: How artificial intelligence can serve mathematical human learning*. Springer, Cham, Switzerland.

Zawacki-Richter, O., Marín, V.I., Bond, M., et al. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education,* 16-39.

II.5
Chatbot e valutazione formativa nella didattica universitaria: evidenze da uno studio pilota di un corso blended

Claudia Baiata
Università degli Studi di Firenze, claudia.baiata@unifi.it
Gabriele Biagini
Università degli Studi di Firenze, gabriele.biagini@unifi.it
Alice Roffi
Università degli Studi di Firenze, alice.roffi@unifi.it
Maria Ranieri
Università degli Studi di Firenze, maria.ranieri@unifi.it

Abstract

Il contributo esplora le potenzialità dell'intelligenza artificiale nella valutazione formativa, integrando feedback automatizzati in un corso universitario in modalità blended. Sono stati analizzati opportunità e limiti di un chatbot progettato e istruito per sostenere l'apprendimento, favorire la partecipazione attiva delle studentesse e supportare la valutazione formativa. Le partecipanti, lavorando in modo collaborativo, hanno riconosciuto l'utilità dell'esperienza e del supporto ricevuto per la progettazione in contesti educativi laboratoriali.

Parole chiave: Valutazione formativa; chatbot; università; blended learning.

This contribution explores the potential of artificial intelligence in formative assessment by integrating automated feedback into a blended university course. The study examined the strengths and limitations of a chatbot designed and instructed to support learning, encourage students' active participation, and enhance formative assessment. Working collaboratively, the participants recognized the value of the experience and the support received for instructional design in laboratory-based educational contexts.

Keywords: formative assessment; chatbot; university; blended learning.

1. Introduzione

L'integrazione dell'intelligenza artificiale (IA) nei processi educativi sta trasformando in modo significativo le pratiche didattiche e valutative, offrendo nuove opportunità per supportare studenti e docenti. In particolare, gli strumenti di feedback automatizzato e i chatbot pedagogici si stanno gradualmente affermando come strumenti funzionali alla valutazione formativa, consentendo di monitorare i progressi, personalizzare il supporto e favorire una partecipazione più attiva e consapevole degli apprendenti. In questo scenario si colloca il progetto qui presentato, realizzato nell'ambito di un percorso universitario magistrale e volto a esplorare l'uso di un chatbot progettato ad hoc per accompagnare le studentesse nella co-progettazione di un MOOC. L'esperienza si inserisce nel più ampio contesto del progetto PRIN *Active Online Assessment*, che mira a comprendere come le tecnologie digitali possano potenziare la valutazione formativa in ambienti online e blended.

2. Riferimenti teorici

Negli ultimi anni, l'impiego degli strumenti di IA in educazione è aumentato in modo significativo, anche in risposta alle diverse complessità che caratterizzano il mondo accademico. In particolare, strumenti di feedback automatico, sistemi di tutoraggio intelligente e soluzioni di *learner profiling* sono stati introdotti con l'obiettivo di supportare i docenti nei processi di valutazione formativa (Bond et al., 2024; Wambsganss, Winkler, Schmid & Söllner, 2020).

Per quanto riguarda l'impiego di tali strumenti da parte dei docenti, un ambito di particolare rilevanza è rappresentato dai processi di valutazione formativa. Si tratta infatti di un settore in cui l'IA può offrire un supporto significativo sia nell'elaborazione di feedback tempestivi e mirati, sia nel monitoraggio continuo dei progressi degli studenti, sostenendo così un insegnamento più attento e in linea con le reali esigenze di apprendimento. L'uso dei sistemi di IA nella valutazione formativa risponde a diverse finalità, coinvolgendo sia la dimensione dello studente, sia quella del docente e della gestione amministrativa. Tra gli obiettivi principali si evidenzia quello di fornire agli studenti un feedback significativo sugli esiti dell'apprendimento in modo individualizzato e personalizzato, al fine di stimolare la riflessione metacognitiva e individuare eventuali misconcezioni (Zhai & Nehm, 2023). Parallelamente, la valutazione formativa offre ai docenti informazioni utili sull'efficacia delle proprie pratiche didattiche, così da consentire un eventuale riorientamento della loro azione educativa, anche in

considerazione del tempo limitato a disposizione per la valutazione degli studenti. L'ausilio dell'IA risulta particolarmente valido per i corsi erogati in modalità online o blended, spesso caratterizzati da un alto numero di studenti e da una notevole eterogeneità, per cui non sempre è possibile avere una conoscenza dei propri studenti, né offrire un'attenzione individualizzata (Puertas, Mariscal-Vivas & Martínez-Requejo, 2023).

La letteratura scientifica che esplora il rapporto tra IA e processi di insegnamento-apprendimento sottolinea l'importanza di un'integrazione consapevole e pedagogicamente orientata (Ranieri & Gaggioli, 2025). Supportando le competenze didattiche dei docenti, gli strumenti *AI enhanced* contribuiscono a migliorare i risultati di apprendimento degli studenti e ad elevare i loro livelli di rendimento. Allo stesso modo, lo studente dovrebbe essere incoraggiato ad assumere un ruolo attivo, attraverso la sua partecipazione responsabile al processo di valutazione formativa, utilizzando le tecnologie come sostegno allo sviluppo di competenze più profonde e non come semplice automatizzazione delle verifiche. Tuttavia, molti autori sottolineano che l'introduzione degli strumenti di IA non può in alcun modo sostituire la funzione educativa del docente, il quale è chiamato a mantenere un ruolo di guida attraverso una mediazione critica, riflessiva e metodologicamente informata, oltre a garantire il suo giudizio imparziale e obiettivo nella valutazione (Ranieri, 2024).

In ambito universitario, i chatbot in particolare stanno conoscendo una certa diffusione come strumenti di supporto per le attività amministrative, la didattica e la valutazione. Tra i loro impieghi più rilevanti vi è l'integrazione del feedback per la valutazione formativa, che consente non solo di monitorare il progresso degli studenti, ma anche di aumentare la motivazione allo studio e migliorare complessivamente i processi di apprendimento (Keyamo, Edun & Baale, 2025). Sviluppati per dialogare con gli utenti attraverso interfacce conversazionali, i chatbot offrono informazioni e suggerimenti in tempo reale; inoltre, comprendono il linguaggio naturale riuscendo ad adattarsi alle esigenze delle persone e rendendo l'esperienza più reale e confortevole. Gli assistenti virtuali possono offrire un supporto concreto, interagendo con spiegazioni ed esempi reali, sia in forma scritta che vocale, e accompagnando gli utenti nelle loro attività. I chatbot possono modificare le proprie domande e risposte in base all'interazione dialogica con gli studenti, favorendo la produzione di feedback più dettagliati (Wambsganss et al., 2020). Grazie alla loro capacità di simulare una conversazione umana, trovano un impiego ideale nei contesti educativi in cui è richiesto un tutor o un feedback automatizzato agli studenti (Puertas, Mariscal-Vivas & Martínez-Requejo, 2023).

Nonostante si configurino come una nuova frontiera in educazione, la lette-

ratura evidenzia anche posizioni contrastanti sull'uso dei chatbot in educazione (Fedeli, Zanon, Pascoletti, D'Agostini & Di Barbora, 2024): da una parte si esprime il timore che possano generare apprendimenti poco accurati e limitare l'interazione umana, dall'altra se ne riconosce il potenziale per sostenere l'autonomia degli studenti. Un uso acritico può incidere sulla valutazione di studenti e studentesse, sull'affidabilità delle informazioni e sullo sviluppo delle competenze di scrittura e creatività. Per questo è essenziale che i docenti siano supportati con un'adeguata formazione che prevenga un utilizzo indiscriminato, così da garantire un'adozione consapevole e vigilante (Bonavolontà & Pagliara, 2024). Un'ulteriore preoccupazione riguarda il rischio di eccessiva dipendenza dagli strumenti di IA, che potrebbe compromettere lo sviluppo del pensiero critico e delle capacità di problem solving degli studenti. Ricerche recenti hanno evidenziato come l'uso frequente di modelli linguistici di grandi dimensioni possa ridurre il carico cognitivo ma, al contempo, limitare la profondità del ragionamento e dell'argomentazione (Stadler, Bannert & Sailer, 2024). Emerge pertanto la necessità di promuovere un utilizzo consapevole di questi strumenti, affinché fungano da supporto al pensiero critico piuttosto che da suo sostituto.

Il presente contributo riguarda l'utilizzo dei chatbot per facilitare la comprensione e la riflessione durante un lavoro collaborativo di macro-progettazione didattica. Tramite il feedback fornito e il supporto immediato agli studenti, è stato possibile, tra l'altro, offrire indicazioni utili per verificare la coerenza tra gli obiettivi formativi e le attività proposte.

3. Contesto

Questo lavoro prende le mosse da una più ampia iniziativa, vale a dire il PRIN *Active Online Assessment in Higher Education* (2023-25) (PRIN - Grant 2022 - Prot. 2022F74XBL) in svolgimento – tra gli altri – negli Atenei di Firenze e Padova. L'obiettivo principale è stato quello di orientare i docenti universitari nella selezione consapevole e strategica di strumenti di valutazione formativa disponibili in accordo con il loro approccio metodologico. Un ulteriore obiettivo è stato verificare in che modo tali strumenti possano migliorare le performance degli studenti e supportare lo sviluppo delle loro competenze metacognitive. Il progetto si è posto anche la finalità di investigare le modalità con cui le tecnologie digitali innovative possano rafforzare le pratiche di valutazione formativa in ambienti online e blended dell'istruzione terziaria.

Per le sperimentazioni svolte è stata adottata la metodologia della Design-Based Research che, con i suoi cicli iterativi e la stretta collaborazione tra ricer-

catori e comunità di pratici, propone una prospettiva che privilegia un intervento immediato, ampliabile e simultaneo nei campi della ricerca, della teoria e dell'applicazione pratica (Wang & Hannafin, 2005). Inoltre, la Design-Based Research contribuisce a favorire l'apprendimento dei partecipanti coinvolti nello studio, grazie all'impegno di produrre risultati che abbiano un impatto tangibile e una ricaduta nel contesto studiato, fornendo una comprensione più approfondita dei suoi meccanismi e delle sue dinamiche (Barab, 2014).

All'interno della sperimentazione condotta in cinque corsi universitari, si è delineata la possibilità di realizzare uno studio pilota in un Corso di Laurea Magistrale caratterizzato da un gruppo ridotto di studentesse partecipanti, erogato in modalità blended learning.

4. Descrizione dell'esperienza

Il contesto e le attività

La sperimentazione di seguito descritta è stata realizzata nel corso "Metodi e tecniche della formazione online", offerto all'interno del Corso in "Scienze pedagogiche e management della formazione per lo sviluppo sostenibile" dell'Università degli Studi di Firenze, strutturato in 22 ore di attività didattiche in presenza e 18 online. Le sei studentesse hanno lavorato alla co-progettazione di un Massive Open Online Course (MOOC) come compito autentico centrale del percorso formativo. L'attività rientrava nella parte applicativa del corso, basata sulla collaborazione tra gruppi. Ogni gruppo doveva definire la macro-progettazione di un MOOC (titolo, destinatari, prerequisiti, carico di lavoro), elaborare il syllabus e organizzare i contenuti in moduli, indicando per ciascuno esiti di apprendimento, attività e modalità di valutazione. Il MOOC previsto era composto da tre moduli, ciascuno articolato in sei unità.

L'inizio di questa attività è avvenuto in presenza per introdurre l'utilizzo degli strumenti di progettazione e di supporto: un'apposita scheda scheda per fornire una guida nella definizione degli elementi della macro-progettazione del MOOC e un chatbot personalizzato per supportare la compilazione della scheda. Una volta acquisita dimestichezza con gli strumenti, le studentesse hanno proseguito l'attività in modalità collaborativa online. Il MOOC co-progettato era focalizzato sul tema della figura dell'Educatore Socio-Pedagogico, da utilizzare come prodotto formativo nella didattica universitaria. La macro-progettazione generale è stata condotta in maniera congiunta dalle studentesse e successivamente la progettazione dei singoli moduli è avvenuta in coppie. Alle studentesse è stato chiesto di realizzare una prima progettazione a cui è stato dato un feed-

back formativo sulla coerenza tra macro-progettazione generale e macro-progettazione dei moduli. In base alle indicazioni fornite, le studentesse hanno apportato le modifiche richieste, rilasciando il prodotto finale, oggetto della restituzione organizzata in presenza per la fine del corso.

Il chatbot come supporto alla progettazione

Per sostenere questo processo è stato introdotto un chatbot basato su GPT, progettato e istruito ad hoc per coadiuvare le discenti nella progettazione e nel progressivo miglioramento del prodotto. Il supporto è stato configurato come un *GPT personalizzato*, al quale sono state fornite istruzioni dettagliate sotto forma di *istruzioni personalizzate* e *template*: oltre agli elementi strutturali della scheda di progettazione, sono stati inclusi criteri di qualità e modelli di riferimento per la formulazione di obiettivi e attività didattiche. Il sistema era finalizzato a guidare passo-passo la progettazione di un MOOC a partire da una scheda di progettazione predefinita che includeva le principali sezioni di macro-progettazione, il syllabus e la struttura modulare del corso. Il sistema ripete lo stesso ciclo di lavoro in ogni fase della progettazione, seguendo istruzioni interne che ne guidano il funzionamento: attraverso una domanda mirata – ad esempio: "Quali obiettivi di apprendimento intendi raggiungere con questo modulo?" oppure "Quante ore di impegno complessivo prevedi per il completamento del MOOC?" – raccoglie le informazioni necessarie sul lavoro specifico; sulla base della risposta dell'utente, fornisce un feedback formativo strutturato sui criteri della scheda di progettazione. Infine, propone la prossima azione – passando alla fase successiva o suggerendo una revisione – e offre un breve riepilogo delle aree di miglioramento emerse nello scambio, come descritto in Figura 1.

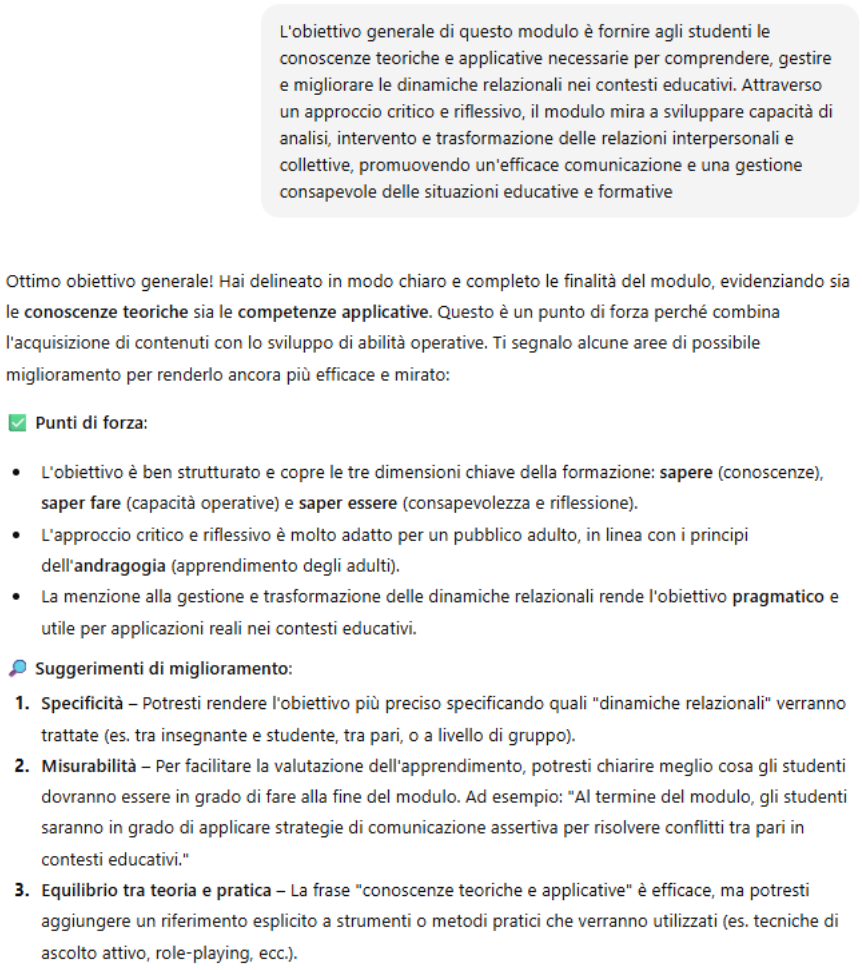


Fig. 1 – Interazioni tra utente e chatbot

Il disegno pedagogico sottostante integra le istanze della progettazione didattica – in particolare l'allineamento costruttivo e la coerenza tra obiettivi formativi, attività e dispositivi valutativi – con i principi dell'andragogia, valorizzando l'autonomia delle studentesse, la rilevanza dei compiti proposti e l'immediata applicabilità delle scelte progettuali. In questa prospettiva, il chatbot non svolge un ruolo meramente correttivo, ma si configura come un mediatore dialogico che sollecita argomentazioni, revisioni e decisioni consapevoli rispetto alla qualità della progettazione. In breve, svolge un ruolo formativo.

Il sistema genera, infatti, un feedback progettato secondo cinque proprietà

chiave: è costruttivo, perché evidenzia sia i punti di forza sia le aree di miglioramento rispetto ai criteri della scheda di progettazione; guidato, grazie a suggerimenti operativi concreti su come riformulare elementi progettuali o rendere più chiari compiti e criteri; motivante, poiché riconosce i progressi e incoraggia ulteriori revisioni; specifico, in quanto fa riferimento diretto ai contenuti proposti dalle studentesse; e pedagogicamente fondato, richiamando esplicitamente principi come l'allineamento tra obiettivi e valutazione, la centralità del *learner* e la chiarezza dei compiti. Dal punto di vista operativo, il sistema fa uso di strategie di *prompting* e *reverse prompting*. Quando l'input risulta incompleto o ambiguo o incoerente, il chatbot chiede chiarimenti, sollecita integrazioni e segnala le eventuali incongruenze. Al termine di ogni sezione della scheda di progettazione, il chatbot rilegge lo scambio e genera due output: una lista di "next steps", che riassume le azioni consigliate (ad esempio chiarire il target o rivedere alcuni *learning outcomes*), e una sintesi metacognitiva che invita a riflettere sul feedback ricevuto e sulle scelte progettuali. Questo trasforma la co-progettazione in un percorso iterativo di miglioramento, in cui il feedback automatizzato funge da valutazione formativa, sostenendo sia la qualità del prodotto finale sia lo sviluppo delle competenze progettuali e riflessive.

Esiti dell'esperienza

Durante la sperimentazione non sono stati previsti *debriefing* collettivi sui feedback del chatbot; i gruppi li hanno elaborati autonomamente in aula. I *log* delle interazioni sono stati raccolti come traccia del processo ma, dato il numero limitato di partecipanti, vengono qui utilizzati solo come materiale esplorativo a supporto della descrizione del setting e delle potenzialità dello strumento. Le opinioni delle studentesse sull'esperienza sono state raccolte attraverso un questionario somministrato al termine dell'attività. L'attività è stata particolarmente apprezzata dalle studentesse per la sua natura pratica e l'utilizzo innovativo di strumenti tecnologici a supporto delle attività: hanno infatti confermato l'utilità del chatbot personalizzato, sia come supporto alla progettazione del MOOC, sia come strumento di verifica della progettazione effettuata. Sono state realizzate delle macro-progettazioni di qualità molto buona, definite coerentemente in ogni aspetto, dagli obiettivi, alle competenze acquisite e alla descrizione degli argomenti trattati in ogni modulo. Tuttavia, la coerenza tra la macro-progettazione generale e la macro-progettazione dei singoli moduli ha rappresentato una criticità durante il processo di progettazione. Di conseguenza, le studentesse hanno particolarmente apprezzato il supporto fornito dal chatbot nel verificare la corrispondenza tra macro-progettazione generale e dei moduli, che ha permesso quindi di evidenziare le parti da migliorare o modificare.

5. Conclusioni

L'esperienza descritta suggerisce come l'impiego di un chatbot progettato per la macro-progettazione di un MOOC possa costituire un valido supporto ai processi di valutazione formativa, favorendo al tempo stesso lo sviluppo di competenze progettuali e riflessive. Il sistema, attraverso feedback strutturati e iterativi, ha contribuito a guidare le studentesse verso una maggiore consapevolezza delle proprie scelte didattiche, sostenendo la coerenza interna del progetto e promuovendo un apprendimento autentico e collaborativo. Pur trattandosi di uno studio pilota, i risultati ottenuti indicano che strumenti di IA, se integrati in modo critico e pedagogicamente orientato, possono rafforzare le pratiche di insegnamento e offrire un valore aggiunto nei contesti universitari.

Lo studio presenta tuttavia alcuni limiti che ne circoscrivono la generalizzabilità. In primo luogo, il numero ridotto di partecipanti e la natura esplorativa dell'indagine non consentono di trarre conclusioni di carattere generale sull'efficacia dello strumento. Inoltre, l'assenza di un gruppo di controllo impedisce di isolare con precisione il contributo specifico del chatbot rispetto ad altre variabili del contesto didattico. Sul piano delle implicazioni, l'esperienza evidenzia come l'integrazione di strumenti di IA nella didattica universitaria richieda una progettazione attenta, che tenga conto sia degli obiettivi formativi sia delle caratteristiche degli apprendenti, evitando il rischio di un utilizzo meramente strumentale o decontestualizzato della tecnologia. Saranno necessari ulteriori approfondimenti per esplorare l'impatto a lungo termine di tali dispositivi e definire modelli di utilizzo sempre più efficaci e consapevoli. In particolare, ricerche future potrebbero indagare l'efficacia di chatbot progettati per la valutazione formativa su campioni più ampi e in contesti disciplinari diversificati, nonché analizzare le dinamiche di interazione tra studenti e sistemi di IA per comprendere meglio come favorire un uso critico e autonomo di questi strumenti.

Riferimenti bibliografici

Barab, S. (2014). Design-based research: A methodological toolkit for engineering change. *The Cambridge handbook of the learning sciences*, *2*, 151-170.

Bonavolontà, G., & Pagliara, S. M. (2024). Tra Timor e Vereor: Il Complesso Dibattito sul Rapporto Uomo-Macchina e la valenza dei Chatbot IA nei contesti educativi. *Nuova Secondaria Ricerca*, *41*(8), 293-303.n

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., ... & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher edu-

cation: A call for increased ethics, collaboration, and rigour. *International journal of educational technology in higher education, 21*(1), 4.

Fedeli, D., Zanon, F., Pascoletti, S., D'Agostini, M., & Di Barbora, E. (2024). Le chatbot: sfida e opportunità per percorsi di didattica metacognitiva. *NUOVA SECONDARIA, 41*(8).

Keyamo, C. A., Edun, P. O., & Baale, A. A. (2025). Development of a Formative Assessment Chatbot for a 100-Level Course in Cybersecurity. *FNAS Journal of Computing and Applications, 2*(3), 62-72.

Puertas, E., Mariscal-Vivas, G., & Martínez-Requejo, S. (2023, November). Development of chatbots connected to Learning Management Systems for the support and formative assessment of students. In *Proceedings of the 2023 7th International Conference on Education and E-Learning* (pp. 14-18).

Ranieri, M. (2024). Intelligenza artificiale a scuola. Una lettura pedagogico-didattica delle sfide e delle opportunità. *Rivista di Scienze dell'Educazione, 62*(1), 123-135.

Ranieri, M., & Gaggioli, C. (2025). *Innovazione didattica e ambienti inclusivi all'università. Dalle competenze digitali all'intelligenza artificiale*. *ETS*.

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior, 160*, 108386.

Wambsganss, T., Winkler, R., Schmid, P. S., & Söllner, M. (2020). Designing a conversational agent as a formative course evaluation tool. In *15th International Conference on Wirtschaftsinformatik*.

Wang, F., & Hannafin, M. J. (2005). Design-Based Research and Technology-Enhanced Learning Environments. *Educational Technology Research and Development*, *53*(4), 5-23.

Zhai, X., & Nehm, R. H. (2023). AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching*, *60*(6), 1390-1398.

II.6
Ricostruire il compito, costruire la rubrica: esperienze di valutazione con il supporto dell'IA

Ilaria Fiore[1]
Università Telematica Pegaso, ilaria.fiore@unipegaso.it
Maria Clara Dicataldo
Università Telematica Pegaso, mariaclara.dicataldo@unipegaso.it
Mariasole Antonietta Guerriero
Università di Foggia, mariasole.guerriero@unifg.it
Alessia Scarinci
Università del Salento, alessia.scarinci@unisalento.it
Anna Dipace
Università Telematica Pegaso, anna.dipace@unipegaso.it

Abstract

L'articolo presenta un laboratorio di Docimologia rivolto a studenti di Scienze della Formazione Primaria dell'Università del Salento, finalizzato allo sviluppo di competenze valutative tramite la progettazione di compiti autentici e rubriche, nell'a.a. 2024/25. Il percorso, fondato su apprendimento attivo e riflessione metacognitiva, integra l'Intelligenza Artificiale generativa (GenAI) come strumento di supporto per ideare, analizzare e migliorare compiti e criteri, stimolando un uso critico e consapevole delle tecnologie. L'esperienza mostra come la GenAI possa facilitare chiarezza, struttura e ricchezza valutativa, mantenendo centrale la mediazione del docente. La ricostruzione di compiti autentici da elaborati reali ha rafforzato la distinzione tra giudizi impressionistici e criteriali. Gli studenti hanno sviluppato competenze professionali emergenti, tra cui prompting efficace e capacità di valutazione trasparente, riconoscendo al contempo limiti e potenzialità della GenAI.

Parole chiave: Valutazione; Compito autentico; Rubriche valutative; Intelligenza Artificiale; Formazione Docenti.

1 Il contributo è frutto del lavoro comune delle autrici. Tuttavia, possono essere attribuiti a Anna Dipace il paragrafo 1; a Maria Clara Dicataldo il paragrafo 2; a Mariasole Antonietta Guerriero il paragrafo 3; ad Alessia Scarinci i paragrafi 4 e 7; a Ilaria Fiore i paragrafi 5 e 6.

The article presents a Docimology laboratory for Primary Education students of the University of Salento to develop assessment competences through the design of authentic tasks and rubrics in the a.y. 2024–25. The workshop, grounded in active learning and metacognitive reflection, integrates generative AI as a support tool to design, analyse, and refine tasks and criteria, fostering a critical and informed use of technology. The experience shows how AI can enhance clarity, structure, and evaluative richness while keeping the teacher's mediation central. Reconstructing authentic tasks from real pupils' work strengthened the distinction between impressionistic and criteria-based judgments. Students developed emerging professional competences, including effective prompting and transparent assessment skills, while recognising both the potential and the limitations of AI.

Keywords: Assessment; Authentic Task; Evaluation Rubrics; Artificial Intelligence; Pre-Service Teacher Training.

1. Oggetto dell'esperienza

Il laboratorio di Docimologia, svolto presso l'Università del Salento con studenti del secondo anno del corso di laurea in Scienze della Formazione Primaria, si colloca all'interno di un percorso più ampio, dedicato alla costruzione di competenze valutative da parte dei futuri insegnanti. La progettazione dell'esperienza nasce dall'esigenza, sempre più rilevata nella letteratura pedagogica e nelle richieste del mondo scolastico, di formare docenti capaci di leggere, documentare e valutare processi di apprendimento complessi, superando logiche trasmissive e classificatorie a favore di una valutazione autentica e orientata allo sviluppo (Darling-Hammond & Snyder, 2000; Trinchero, 2023).

L'esperienza laboratoriale è stata progettata per offrire agli studenti un ambiente di apprendimento situato e partecipativo, in cui sperimentare in modo diretto la costruzione di compiti autentici e rubriche valutative. La scelta di introdurre l'Intelligenza Artificiale Generativa (GenAI) come strumento di supporto si colloca all'interno del dibattito contemporaneo sulle possibilità offerte dalla GenAI nei processi educativi, con l'obiettivo non solo di presentare strumenti operativi, ma anche di promuovere una prospettiva critica sul loro utilizzo. Il laboratorio, infatti, ha inteso far emergere sia il potenziale della GenAI nel facilitare la progettazione valutativa sia le tensioni metodologiche, etiche e pedagogiche che essa introduce nel lavoro dell'insegnante.

Gli obiettivi formativi sono stati quindi molteplici:

– comprendere la valutazione delle competenze e la natura dei compiti autentici;
– conoscere e costruire rubriche valutative coerenti, affidabili e trasparenti;
– utilizzare GenAI come co-progettista, imparando a riconoscerne punti di forza e criticità;
– riflettere sui processi di *feedback* e sulla distinzione tra giudizio impressionistico e valutazione criteriale;
– promuovere *agency*, collaborazione e consapevolezza metacognitiva tra gli studenti coinvolti.

La dimensione esperienziale, la ciclicità delle attività e la continua alternanza tra teoria e pratica hanno permesso agli studenti di agire come progettisti e valutatori, costruendo progressivamente una competenza riflessiva e professionale che si intreccia con le sfide della scuola contemporanea.

La cornice teorica dell'esperienza si articola attorno a tre nuclei concettuali: la valutazione per competenze e i compiti autentici, le rubriche come strumenti di trasparenza e affidabilità della valutazione, e il ruolo della GenAI nei processi valutativi.

2. Valutazione per competenze e rubriche valutative

Negli ultimi anni la valutazione per competenze ha assunto un ruolo centrale nelle politiche e nelle pratiche educative, configurandosi come un paradigma orientato alla mobilitazione integrata di conoscenze, abilità e atteggiamenti in situazioni significative (Mannonen et al., 2025).

Tale approccio richiede strumenti capaci di cogliere la complessità dell'agire dello studente, andando oltre la verifica di contenuti isolati (Piricò et al., 2023). In questo quadro il compito autentico si configura come un dispositivo privilegiato: esso sollecita lo studente a risolvere problemi o affrontare situazioni vicine alla realtà, attivando processi di trasferibilità e competenze trasversali (Tessaro, 2014).

La letteratura internazionale, come evidenziato anche da Vlachopoulos e Makri (2024), mostra come l'*authentic assessment* favorisca lo sviluppo delle abilità del XXI secolo, incrementi la motivazione, sostenga processi di *engagement* e avvicini il curricolo universitario alle competenze richieste nei contesti professionali. Inoltre, la natura situata e contestualizzata dei compiti autentici facilita la valutazione del processo oltre che del prodotto, offrendo informazioni più ricche e formative per l'apprendimento.

Le rubriche rappresentano uno strumento chiave per rendere espliciti e con-

divisi i criteri di valutazione, aumentando coerenza, equità e trasparenza. Esse permettono allo studente di orientarsi nella realizzazione del compito e di esercitare processi di autoregolazione, autovalutazione e *peer-feedback* (Panadero et al., 2023). Studi recenti (Bin Dahmash, 2025; Panadero et al., 2025) confermano che l'uso sistematico delle rubriche incrementa l'affidabilità inter-valutatore, potenzia le abilità metacognitive e migliora la qualità dei *feedback*, soprattutto quando integrate in percorsi formativi che prevedono la co-costruzione o la discussione dei criteri.

Tuttavia, la letteratura segnala anche alcune criticità: la necessità di adeguata formazione dei docenti, il rischio di un'eccessiva analiticità che irrigidisce la valutazione, la possibile difficoltà di mantenere rubriche coerenti con la complessità dei compiti autentici (Panadero & Jönsson, 2020; Ling, 2024). In questo senso, il lavoro con gli studenti universitari rappresenta un'occasione per sviluppare competenze professionali consapevoli e orientate alla qualità (Isusi-Fagoaga & García-Aracil, 2020).

3. GenAI come supporto alla didattica

L'integrazione della GenAI nei processi valutativi apre scenari promettenti ma anche problematici. Le ricerche più recenti (Prompiengchai et al., 2025; Usher, 2025) evidenziano che la GenAI possa sostenere la progettazione di compiti e rubriche, offrire feedback rapidi e personalizzati, facilitare il monitoraggio continuo e favorire l'autoregolazione degli studenti. La sua capacità di generare esemplificazioni, revisioni linguistiche, *check-list* o alternative progettuali può rappresentare un valore aggiunto nel lavoro del docente, soprattutto nella fase di ideazione.

Al tempo stesso emergono rischi legati a fenomeni di disallineamento, allucinazioni, sovrastima dell'accuratezza, perdita di *agency* e possibili distorsioni etiche (Kalonde at al., 2025). L'IA non può sostituire la capacità interpretativa del docente, né il valore relazionale e dialogico del *feedback* formativo. La riflessione critica sul ruolo della GenAI diventa pertanto un obiettivo educativo primario: solo sviluppando competenze valutative e digitali integrate, i futuri insegnanti possono utilizzare tali strumenti in modo consapevole, mantenendo il controllo epistemico e professionale dei processi di valutazione.

In sintesi, la combinazione tra compiti autentici, rubriche e GenAI richiede un ripensamento della valutazione che metta al centro processi, riflessività e trasparenza, superando l'idea della valutazione come semplice misurazione o controllo dell'originalità. Il laboratorio presentato nel contributo si colloca esattamente in questa prospettiva trasformativa.

4. Contesto e partecipanti

Il laboratorio di Docimologia ha coinvolto circa 100 studenti del secondo anno di Scienze della Formazione Primaria e si è svolto nel mese di maggio dell'anno accademico 2024-2025, in due incontri di tre ore l'uno, alternando momenti frontali, attività di gruppo e riflessioni guidate con strumenti digitali. La suddivisione in gruppi e le restituzioni plenarie hanno favorito collaborazione, discussione e consolidamento delle pratiche valutative. In linea con la letteratura sulla didattica laboratoriale in ambito universitario, l'impianto metodologico ha promosso un apprendimento attivo e situato, nel quale gli studenti costruiscono conoscenze attraverso compiti autentici, interazione tra pari e riflessione guidata (Freeman et al., 2014). Tale approccio, fondato su un ciclo di esperienza, rielaborazione e applicazione, tipico dell'apprendimento esperienziale (Kolb, 1984), favorisce lo sviluppo di competenze professionali e metacognitive rilevanti per la formazione dei futuri docenti.

5. Descrizione dell'attività svolta

Il laboratorio si è svolto in due incontri di tre ore, con una metodologia laboratoriale. Nello specifico, il percorso si è articolato nelle seguenti fasi di lavoro:

- avvio e introduzione concettuale con un *ice-breaker* didattico;
- *task* di gruppo di progettazione di compiti autentici e rubriche con e senza strumenti GenAI;
- *task* di ricostruzione e valutazione di compiti autentici, confrontando valutazioni intuitive con valutazioni criterali anche mediante il supporto di strumenti di GenAI;
- feedback e valutazione riflessiva sull'esperienza condotta all'interno dell'ambiente digitale *Wakelet*.

Nella prima fase, dopo un momento di *ice-breaking* mediante l'applicazione *Wooclap*[2] sul concetto di valutazione, con l'obiettivo di stimolare consapevolezza metacognitiva e collegamento con la costruzione delle rubriche (Figura 1), gli studenti hanno lavorato sulla progettazione di compiti autentici e rubriche, a partire dai contenuti teorici esposti dal docente.

2 https://www.wooclap.com/it/ (ultimo accesso 10–02-2026).

In un primo momento hanno lavorato senza il supporto della GenAI, confrontandosi tra pari sulle scelte da compiere e sui criteri da adottare. In questa fase il *task* prevedeva l'ideazione di un compito autentico rivolto a studenti della scuola primaria, basato sulla realizzazione di un prodotto concreto o sulla risoluzione di un problema in un contesto realistico (es. scrivere una lettera, progettare un cartellone informativo, realizzare un esperimento semplice, ecc.). Agli studenti è stato chiesto di descrivere il compito, specificando: obiettivi, destinatari, contesto, prodotto atteso (Figura 2) e di costruire una rubrica di valutazione collegata al compito, contenente almeno 3 criteri di valutazione e 3 livelli di prestazione (Figura 3).

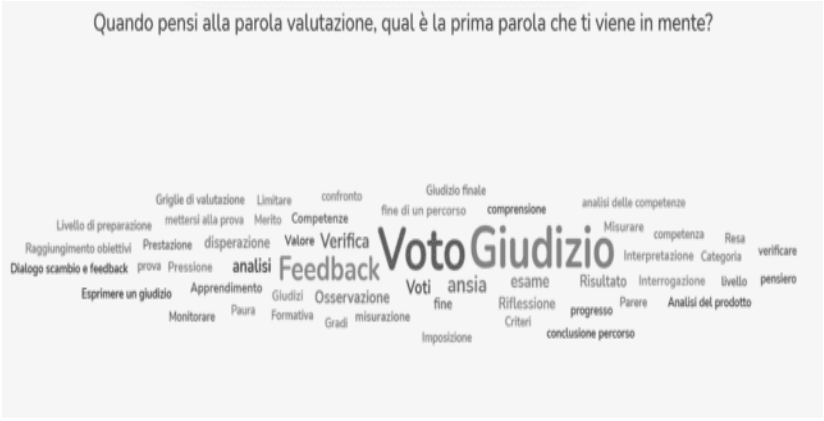


Fig. 1 – Word cloud composta dalle parole scelte dagli studenti per identificare il costrutto "valutazione"

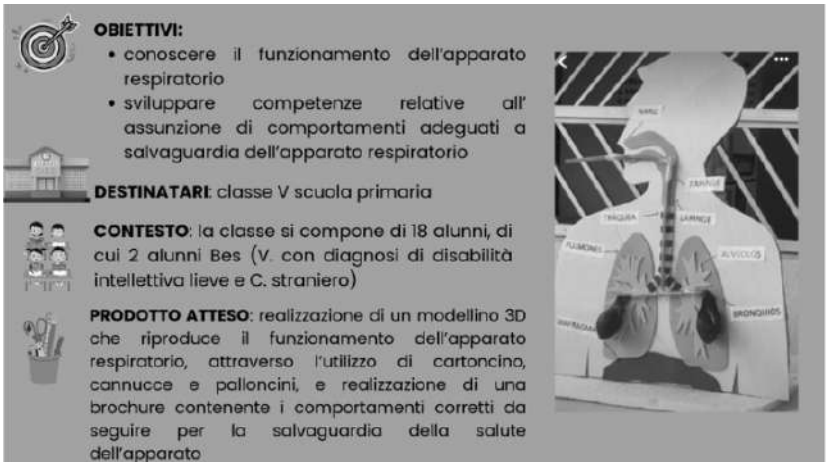


Fig. 2 – Esempio di compito autentico elaborato dai corsisti

RUBRICA DI VALUTAZIONE PRODOTTO				
CRITERI	DESCRITTORI			
	PIENAMENTE RAGGIUNTO	RAGGIUNTO	PARZIALMENTE RAGGIUNTO	IN FASE DI ACQUISIZIONE
COMPLETEZZA DEL MODELLO REALIZZATO	Il modellino risulta realizzato in modo completo in ogni sua parte	Il modellino risulta abbastanza completo in ogni sua parte	Il modellino risulta parzialmente completo	Il modellino risulta incompleto
PRECISIONE NELLA REALIZZAZIONE DEL MODELLO	Il modellino realizzato risulta preciso in ogni sua parte	Il modellino realizzato risulta abbastanza preciso	Il modellino realizzato risulta preciso solo in parte	Il modellino realizzato risulta poco preciso
GESTIONE DEL TEMPO	Gestisce bene il tempo disponibile e rispetta sempre le scadenze stabilite.	Gestisce sufficientemente bene il tempo disponibile e nel complesso rispetta le scadenze stabilite.	Tende a lavorare lentamente, ma, se opportunamente sollecitato rispetta tempi e scadenze stabilite.	Raramente rispetta i tempi stabiliti e le scadenze
FUNZIONALITA' DEL MODELLO	Il modellino proposto è perfettamente funzionante	Il modellino proposto funziona abbastanza bene	Il modellino proposto funziona sebbene con qualche difficoltà	Il modellino proposto non funziona
ATTEGGIAMENTO NEI CONFRONTI DEGLI ALTRI	Ascolta e rispetta gli altri mostrando interesse per le loro ideee e opinioni. Ha un ottimo senso del gruppo e si relaziona positivamente con gli altri.	Di solito ascolta e rispetta gli altri. Ha un buon senso del gruppo e si relaziona positivamente con i compagni all'interno del gruppo.	Tendenzialmente ascolta e rispetta gli altri ma spesso partecipa e interagisce con i compagni solo se sollecitato	Raramente ascolta gli altri e tende ad isolarsi partecipando solo se sollecitato e supportato

Fig. 3 – Rubrica di valutazione elaborata dagli studenti senza il supporto della GenAI

Successivamente, i corsisti si sono cimentati in un laboratorio di revisione del proprio lavoro, adottando gli strumenti di GenAI, *ChatGPT* e *Magic School*[3], per la costruzione delle rubriche valutative. Agli studenti è stato fornito un modello di *prompt* da utilizzare, secondo la tecnica PARTS (Wei *et al.*, 2021): "Sei un insegnante esperto e abile nel creare valutazioni e nel valutare il lavoro degli alunni. Il tuo compito è creare una rubrica per la mia classe di [LIVELLO e MATERIA] che studia [ARGOMENTO]. I miei studenti stanno completando [TITOLO DEL COMPITO], in cui [DESCRIZIONE DEL COMPITO]. Formatta la rubrica come un grafico e includi una scala a [NUMERO] punti".

Tale comando è stato poi applicato dai corsisti, adottando le opportune modifiche, in base al compito autentico sviluppato.

L'introduzione progressiva della GenAI ha permesso agli studenti di confrontare soluzioni umane e soluzioni generate artificialmente, riflettendo sull'accuratezza, sulla trasferibilità e sui limiti dei modelli linguistici nell'ambito valutativo.

3 https://www.magicschool.ai/ (ultimo accesso 10-02-2026).

Criteri di Valutazione	4 - Eccellente	3 - Buono	2 - Sufficiente	1 - Insufficiente
Conoscenza del funzionamento dell'apparato respiratorio	Descrive con precisione e completezza il funzionamento dell'apparato respiratorio utilizzando termini corretti e dettagliati.	Spiega il funzionamento dell'apparato respiratorio in modo chiaro, con qualche dettaglio mancante o imprecisione lieve.	Mostra una conoscenza di base del funzionamento, ma con spiegazioni generiche o incomplete.	Non riesce a descrivere correttamente il funzionamento dell'apparato respiratorio.
Realizzazione del modellino	Costruisce un modellino funzionale, ben curato e dettagliato che rappresenta chiaramente il funzionamento dell'apparato respiratorio.	Realizza un modellino comprensibile e abbastanza dettagliato, con qualche imperfezione o mancanza.	Il modellino è semplice e poco dettagliato, con alcune parti non chiaramente rappresentate.	Il modellino è incompleto o non rappresenta adeguatamente il funzionamento dell'apparato.
Riflessione sugli stili di vita e comportamenti	Identifica e spiega in modo approfondito le abitudini e gli stili di vita che influenzano positivamente o negativamente l'apparato respiratorio.	Riconosce alcune abitudini e stili di vita influenti, con spiegazioni semplici ma pertinenti.	Menziona alcune abitudini, ma con poca chiarezza o collegamento al funzionamento dell'apparato.	Non comprende o non riesce a riflettere sulle abitudini che influenzano l'apparato respiratorio.
Presentazione al gruppo	Presenta con sicurezza, chiarezza e completezza il modellino e le riflessioni, coinvolgendo il gruppo.	Presenta in modo chiaro il modellino e le riflessioni, con qualche esitazione o dettaglio mancante.	Presenta in modo semplice, con poca sicurezza e poca partecipazione del gruppo.	Non presenta o la presentazione è confusa e poco pertinente.

Fig. 4 – Rubrica di valutazione elaborata dagli studenti con il supporto di *MagicSchool*

Nella terza fase gli studenti hanno osservato e ricostruito un compito autentico, partendo da elaborati reali di bambini di scuola primaria, forniti dal docente del corso. Ciascun gruppo ha selezionato un elaborato tra: modellini rappresentanti le civiltà storiche, *lapbook* su costrutti grammaticali di lingua italiana, mappe concettuali su concetti scientifici, manufatti artistici, *task* matematici realistici e descrizioni di luoghi familiari.

Ogni gruppo ha elaborato una traccia, fornito un primo *feedback* libero e successivamente uno critico sui lavori osservati, basato su rubriche, sperimentando nuovamente l'uso della GenAI. Tale attività ha sostenuto la capacità degli studenti di passare da una valutazione impressionistica a una valutazione criteriale, in linea con le funzioni progettuali, orientative e valutative delle rubriche.

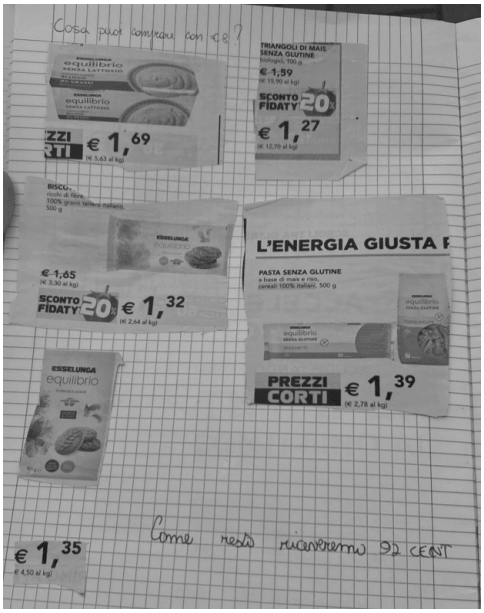


Fig. 5 – Task matematico realizzato da alunni di scuola primaria

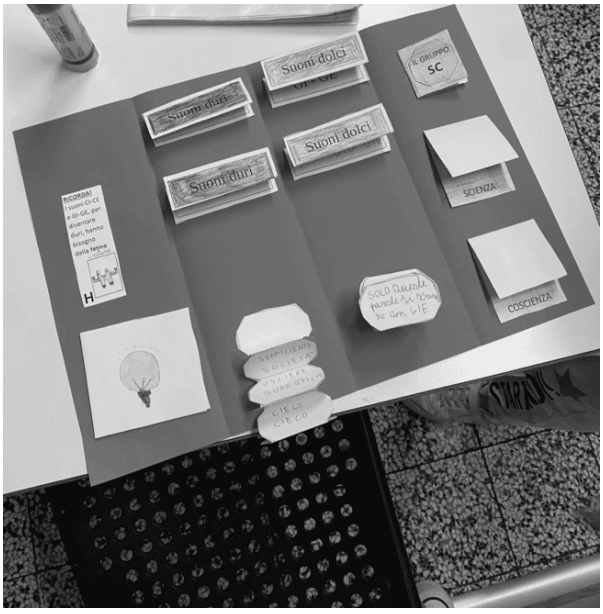
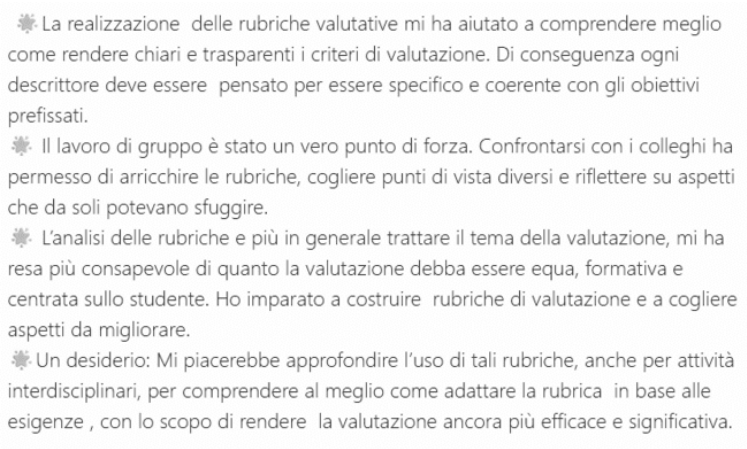


Fig. 6 – Lapbook realizzato da alunni di scuola primaria

Il percorso si è concluso con un compito riflessivo: ogni studente ha scritto un breve post sulla bacheca condivisa digitale, *Wakelet*[4], applicando la strategia "3 stelle e un desiderio" (Petty, 2009), evidenziando gli aspetti più apprezzati e i desideri di miglioramento. Ciò ha contribuito all'integrazione della dimensione soggettiva della valutazione, connessa alla consapevolezza del proprio apprendimento.


La realizzazione delle rubriche valutative mi ha aiutato a comprendere meglio come rendere chiari e trasparenti i criteri di valutazione. Di conseguenza ogni descrittore deve essere pensato per essere specifico e coerente con gli obiettivi prefissati.

Il lavoro di gruppo è stato un vero punto di forza. Confrontarsi con i colleghi ha permesso di arricchire le rubriche, cogliere punti di vista diversi e riflettere su aspetti che da soli potevano sfuggire.

L'analisi delle rubriche e più in generale trattare il tema della valutazione, mi ha resa più consapevole di quanto la valutazione debba essere equa, formativa e centrata sullo studente. Ho imparato a costruire rubriche di valutazione e a cogliere aspetti da migliorare.

Un desiderio: Mi piacerebbe approfondire l'uso di tali rubriche, anche per attività interdisciplinari, per comprendere al meglio come adattare la rubrica in base alle esigenze , con lo scopo di rendere la valutazione ancora più efficace e significativa.

Fig. 7 – Post riflessivo di un corsista

6. Discussione

Le attività, organizzate in gruppi, hanno reso gli studenti protagonisti della progettazione e della sperimentazione di rubriche, sostenendo riflessioni metacognitive guidate da *prompt* strutturati.

In particolare, il *task* di integrazione di strumenti di GenAI per la costruzione delle rubriche valutative ha condotto verso un confronto significativo: mentre le rubriche progettate dagli studenti erano centrate sul prodotto, sulla precisione del modellino e sulla cooperazione, quelle generate dagli strumenti utilizzati, come *ChatGPT* e *Magic School*, risultavano focalizzate soprattutto sulle conoscenze disciplinari. Questo passaggio ha reso possibile un confronto diretto tra le soluzioni elaborate dagli studenti e quelle proposte dalla GenAI, aprendo uno spazio di riflessione sulle differenze emerse nei criteri utilizzati, nella struttura delle rubriche e nel modo di intendere il *focus* della valutazione.

4 https://wakelet.com/wake/uH4XM61O64Vmf276HFNcQ (ultimo accesso 07-12-2025).

La ricostruzione di compiti autentici a partire da elaborati reali di alunni e il confronto tra un feedback libero e un feedback basato su rubriche, invece, hanno consentito una riflessione sul suo ruolo e sul suo valore formativo. Gli studenti hanno sperimentato il passaggio da un feedback spontaneo e intuitivo ad uno più strutturato, fondato su criteri espliciti e rubriche condivise. A partire dagli esempi prodotti dagli studenti è stato possibile mettere in luce come il feedback possa accompagnare lo studente nel miglioramento del proprio lavoro e rendere più visibili i processi di apprendimento. Anche in questa fase, la GenAI è stata utilizzata in modo critico e consapevole, come supporto al processo valutativo e non come sostituto del giudizio professionale del docente.

Tale esperienza, dunque, rientra in una cornice teorica più ampia, che riconosce alla GenAI un ruolo di supporto ai processi professionali, a condizione che sia inserita in contesti formativi autentici e mediata criticamente dal docente. In questa prospettiva, queste riflessioni si allineano con l'idea che la GenAI possa arricchire le capacità progettuali del docente senza sostituirne la professionalità. Come sostiene Luckin (2018), la GenAI può potenziare l'azione educativa, lasciando tuttavia un ruolo centrale al docente, chiamato a mediare, contestualizzare e personalizzare le proposte generate.

Durante le attività e le restituzioni e nei confronti conclusivi con gli studenti si sono potute individuare delle percezioni comuni riguardo alla GenAI in ambito didattico, tra cui:

1. *La GenAI è percepita come un valido strumento di supporto progettuale* in grado di migliorare chiarezza, struttura e ricchezza delle rubriche e dei compiti autentici.
2. *Il ruolo centrale del docente* sia per mediare criticamente, sia per contestualizzare e personalizzare le proposte fornite dalla GenAI.
3. *L'interazione della GenAI come fattore di sviluppo di competenze professionali emergenti* (es. competenza nel prompting per la generazione di valutazioni trasparenti).
4. *La GenAI come stimolo ad una riflessione più ampia sulla valutazione,* promuovendo oggettività, completezza e approfondimento dei criteri valutativi.

Si tratta, dunque, di percezioni che potrebbero rappresentare un terreno per future linee di ricerca.

7. Conclusioni

Nel complesso, l'esperienza può essere interpretata come un contesto esplorativo in cui gli studenti hanno avuto l'opportunità di avvicinarsi a una prima forma di *AI literacy* in ambito professionale, riconoscendone alcune potenzialità (ad esempio in termini di supporto alla chiarezza e all'articolazione dei criteri valutativi) e alcuni limiti (quali la genericità delle risposte, la ridotta contestualizzazione e la complessità linguistica). In questa prospettiva, l'uso dell'IA sembra richiedere l'integrazione con attività metacognitive che rendano esplicito il processo di revisione critica, affinché tali strumenti possano assumere una funzione effettivamente formativa.

Tali considerazioni risultano coerenti con la letteratura sulla formazione iniziale dei docenti, che evidenzia come lo sviluppo di competenze valutative e riflessive richieda dispositivi formativi intenzionalmente progettati, capaci di integrare dimensioni teoriche, operative e riflessive. In particolare, gli studi sulla *assessment literacy* sottolineano come la capacità di progettare, interpretare e utilizzare strumenti valutativi non possa essere data per acquisita, ma debba essere oggetto di una formazione esplicita e progressiva nei percorsi di *Initial Teacher Education* (DeLuca et al., 2016; Popham, 2018).

In tale quadro, l'introduzione di strumenti digitali e di intelligenza artificiale nella formazione iniziale non può essere considerata di per sé sufficiente, ma andrebbe collocata all'interno di percorsi che promuovano un uso critico, consapevole e pedagogicamente mediato delle tecnologie, in linea con le competenze professionali richieste alla docenza contemporanea (OECD, 2021; Redecker, 2017). L'esperienza descritta suggerisce pertanto la necessità di ulteriori ricerche, basate su disegni metodologici più strutturati, per esplorare in modo sistematico il contributo dell'IA allo sviluppo delle competenze valutative nei futuri insegnanti.

Riferimenti bibliografici

Bin Dahmash, N. F. (2025). The analytic use of rubrics in writing classes by language students in an EFL context: Students' writing model and benefits. *Frontiers in Education, 1*,15880460. https://doi.org/10.3389/feduc.2025.1588046

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and teacher education*, 16(5-6), 523-545

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, 28, 251-272.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Isusi-Fagoaga, R., & García-Aracil, A. (2020). Assessing master students' competencies using rubrics: Lessons learned from future secondary education teachers. *Sustainability, 12*(23), 9826. https://doi.org/10.3390/su12239826

Kalonde, G., Boateng, S., & Duedu, C. (2025). Artificial Intelligence in Classroom Assessment: Opportunities, Equity Challenges, and Best Practices for Formative and Summative Integration. *Open Access Library Journal, 12*, 1-16. doi: 10.4236/oalib.1114121

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice-Hall

Ling, J. H. (2024). A Review of Rubrics in Education: Potential and Challenges. Pedagogy: *Indonesian Journal of Teaching and Learning Research, 2*(1), 1-14. https://ejournal.aecindonesia.org/index.php/pedagogy/article/view/199

Luckin, R. (2018). *Machine learning and human intelligence: The future of education for the 21st century*. UCL Institute of Education Press

Mannonen, J., Hooshyar, D., Urrutia, F., Hämäläinen, R., Araya, R., & Lehesvuori, S. (2025). The concept of competence in educational research: a knowledge map of main research traditions and topics. *Education+ Training*, 67(10), 132-149

OECD. (2021). Teachers at the centre of education recovery. OECD Publishing

Panadero, E., & Jönsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review, 30*. https://doi.org/10.1016/j.edurev.2020.100329

Panadero, E., Delgado, P., Zamorano, D., Pinedo, L., Fernández-Ortube, A., & Barrenetxea-Mínguez, L. (2025). Putting excellence first: How rubric performance level order and feedback type influence students' reading patterns and task performance. *Learning and Instruction, 99*, 102168. https://doi.org/10.1016/j.learninstruc.2025.102168

Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review*, 35(4), 113.

Petty, G. (2009). *Evidence-based teaching: A practical approach* (2nd ed.). Cheltenham: Nelson Thornes

Piricò, M. L., Salvisberg, M., & Giovannini, V. (2023). *La valutazione per competenze. Dalla teoria alla prassi*. Repubblica e Cantone Ticino: DECS

Popham, W. J. (2018). *Classroom assessment: What teachers need to know* (8th ed.). Pearson.

Prompiengchai, S., Narreddy, C., & Joordens, S. (2025). A Practical Guide for Supporting Formative Assessment and Feedback Using Generative AI (arXiv:2505.23405). *arXiv*. https://doi.org/10.48550/arXiv.2505.23405

Redecker, C. (2017). E*uropean framework for the digital competence of educators: DigCompEdu*. Publications Office of the European Union.

Tessaro, F. (2014). Compiti autentici o prove di realtà? *Formazione & insegnamento, 12*(3), 77-88. Retrieved from https://ojs.pensamultimedia.it/index.php/siref/article/view/1119

Vlachopoulos, D., & Makri, A. (2024). A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations. *Studies in Educational Evaluation, 83*, 101425. https://doi.org/10.1016/j.stueduc.2024.101425

Trinchero, R. (2023). Assessment as learning in università.: Costruire le capacità autovalutative degli studenti. *Pedagogia oggi*, 21(1), 108-117. https://doi.org/10.7346-/PO-012023-12

Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education, 50*(6), 912–927. https://doi.org/10.1080/02602938.2025.2487495

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). *Finetuned language models are zero-shot learners*. arXiv preprint arXiv:2109.01652

II.7
From Challenge to Opportunity: Critical Use of GenAI in Mathematical Foundations for Architecture Students

Maria Giulia Ballatore[1]
Dept. of Mathematical Sciences 'G.L. Lagrange', Politecnico di Torino

Abstract

This paper presents an exploratory teaching experience integrating Generative AI (GenAI) into mathematical foundations courses for architecture students. The intervention transformed GenAI from a perceived threat into a pedagogical opportunity through a "discuss the screen" methodology implemented across four sessions during a semester. Students developed critical thinking skills by comparing their solutions with GenAI outputs, addressing four pedagogical dimensions: disciplinary (improving mathematical understanding), digital (promoting conscious GenAI use), andragogical (enhancing motivation), and metacognitive (developing transversal reflection skills). Observations from final exams and classroom interactions suggest increased metacognitive awareness, evidenced by more explanatory comments and sustained engagement. This exploratory experience demonstrates how thoughtful integration of GenAI can enhance relational understanding of mathematics while developing critical thinking competencies, offering initial insights that warrant further systematic investigation.

Keywords: Generative Artificial Intelligence; Mathematics Education; Critical Thinking; Metacognition; Architecture students.

Questo contributo presenta un'esperienza didattica esplorativa di integrazione dell'Intelligenza Artificiale Generativa (GenAI) nei corsi di fondamenti matematici per studenti di architettura. L'intervento ha trasformato la GenAI da potenziale minaccia percepita a opportunità pedagogica attraverso una metodologia di "discuss the screen", implementata in quattro sessioni nel corso di un semestre. Gli studenti hanno sviluppato capacità di pensiero critico confrontando le pro-

1 She is a Research Fellow at the Department of Mathematical Sciences of the Politecnico di Torino, Italy. She received her PhD in Electrical, Electronics, and Communications Engineering, focusing on Engineering Education at Politecnico di Torino, Italy. Her research interests lie in the fields of Engineering Education, Generative AI in Higher Education, Spatial Abilities, Playful Learning, and Gender Issues. ORCID: 0000-0002-6216-8939

prie soluzioni con gli output della GenAI, lavorando su quattro dimensioni pedagogiche: disciplinare (miglioramento della comprensione matematica), digitale (promozione di un uso consapevole della GenAI), andragogica (potenziamento della motivazione), e metacognitiva (sviluppo di competenze trasversali di riflessione). Le osservazioni emerse dagli esami finali e dalle interazioni in aula suggeriscono un aumento della consapevolezza metacognitiva, evidenziato da un maggior numero di commenti esplicativi e da un impegno sostenuto nel corso. Questa esperienza esplorativa mostra come un'integrazione critica e intenzionale della GenAI possa favorire una comprensione relazionale della matematica e lo sviluppo di competenze di pensiero critico, offrendo prime evidenze che richiedono ulteriori indagini sistematiche.

Parole Chiave: Intelligenza Artificiale Generativa; didattica della matematica; pensiero critico; metacognizione; studenti di architettura.

1. Introduction

The advent of Generative Artificial Intelligence (GenAI) has profoundly altered the educational landscape, particularly in foundational mathematics courses. Tools like ChatGPT and Claude can solve mathematical exercises instantaneously, creating an illusion that computational skills can be outsourced without understanding the underlying processes. This phenomenon is especially critical in non-characterizing courses, such as mathematical foundations for architecture students, where the perception of utility is already fragile.

At Politecnico di Torino, the course "Istituzioni di Matematiche" (Mathematical Foundations) represents a mandatory 8 ECTS requirement for first-year architecture students. Traditionally perceived as an obstacle rather than an opportunity, this course faces the additional challenge of students questioning its relevance in an era where AI can perform calculations. However, this challenge can be transformed into a pedagogical opportunity by critically integrating GenAI into the learning process.

This paper presents an exploratory educational intervention that reframes GenAI from a tool that threatens to make mathematical skills obsolete into a pedagogical instrument that enhances critical thinking, metacognitive awareness, and relational understanding of mathematics. The approach is grounded in Skemp's distinction between instrumental and relational understanding (Skemp, 1976), combined with scaffolding theory (Wood, Bruner, & Ross, 1976) and a four-dimensional pedagogical framework addressing disciplinary, digital, andragogical, and metacognitive competencies.

This work documents a teaching experience rather than a controlled empirical study. The observations presented emerge from classroom practice and qualitative assessment of student engagement and exam performance. While these initial findings are promising, they represent preliminary evidence requiring further systematic investigation with rigorous research methodology.

2. Theoretical Framework

Skemp (1976) distinguished between two types of mathematical understanding: instrumental understanding ("knowing how to do something") and relational understanding ("knowing both what to do and why"). In traditional mathematics education, particularly in service courses for non-mathematical programs, there is often an overemphasis on instrumental understanding, where students learn algorithms and procedures without grasping the conceptual foundations.

GenAI tools excel at instrumental tasks but fundamentally lack relational understanding. This creates a pedagogical paradox: if we teach mathematics purely instrumentally, we are teaching students to do what machines can already do better. However, if we use GenAI's limitations as a teaching tool, we can highlight the irreplaceable value of relational understanding.

The pedagogical approach draws significantly on Vygotsky's concept of scaffolding and the Zone of Proximal Development (ZPD). Scaffolding refers to the temporary support structures that enable learners to accomplish tasks they cannot yet perform independently (Wood, Bruner & Ross, 1976). In traditional mathematics education, scaffolding typically involves instructor guidance, worked examples, or peer collaboration.

The "discuss the screen" methodology introduces a novel form of scaffolding: using GenAI's imperfect outputs as cognitive scaffolds. Unlike traditional scaffolding where the support is more knowledgeable, GenAI provides a different kind of support: a comparator that forces explicit articulation of reasoning. Students work within their ZPD as they evaluate outputs that are sometimes correct, sometimes subtly wrong, requiring them to stretch beyond their current independent capability while remaining within reach through collaborative discussion. This approach aligns with Wood et al.'s (1976) scaffolding functions: reducing degrees of freedom by focusing attention on specific aspects of solutions, maintaining direction toward the goal, highlighting critical features, and controlling frustration. The gradual release of responsibility, with progressive fading of instructional support as learners gain competence, is characteristic of

effective scaffolding in cognitive apprenticeship models (Collins, Brown & Newman, 1989).

Secondly, the metacognition (defined as the awareness and regulation of one's own thinking processes) is crucial for deep mathematical learning (Schoenfeld, 1992). When students must articulate their reasoning, identify errors, and justify their solutions, they develop metacognitive skills that transcend specific mathematical content. The presence of GenAI creates authentic situations requiring students to explain their thinking explicitly.

3. Context and Methodology

The intervention took place during the 2024/25 academic year in "Istituzioni di Matematiche", a first-semester course for architecture students at Politecnico di Torino. The course comprised approximately 150 students with heterogeneous mathematical backgrounds, typical of architecture programs. The course content includes foundational mathematical topics often perceived as disconnected from architectural practice such as linear algebra, differential calculus, and integral calculus. The course follows a traditional structure with lectures delivered by the course holder (Prof. De Gregorio) and exercise sessions conducted by the author.

The pedagogical intervention centered on a methodology termed "discuss the screen", implemented during exercise sessions. Over 12 exercise sessions throughout the semester, four sessions incorporated this approach (approximately one every three sessions). The methodology follows a structured cycle:

Parallel Problem-Solving: Students receive an exercise to solve individually or in small groups with nearby classmates. Simultaneously, the same exercise is input into GenAI tools (primarily the free version of ChatGPT, with occasional verification using the paid version of Claude).

Individual/Collaborative Work: Students work on the problem for several minutes, applying their understanding and discussing approaches with peers.

Comparative Presentation: The instructor projects the GenAI solution on screen, initiating collective analysis. Students compare their solutions with the GenAI output.

Critical Discussion: When discrepancies emerge, the class engages in dialogue to identify who is correct and why. The instructor facilitates discussion without immediately revealing the answer, encouraging students to justify their reasoning.

Verification and Reflection: The class collectively decides on a follow-up

prompt to deepen the analysis. Critical reading continues iteratively until reaching consensus between student solutions and AI outputs. In some cases, the same exercise is presented using multiple LLM, repeating the cycle from point 3 to stimulate comparative analysis across different AI tools.

Exercises were deliberately chosen from standard course topics: determinants of 4×4 matrices, rank calculations for 4×3 matrices, solutions to linear systems, limits, and calculation of areas between curves. These exercises are procedurally complex enough that both students and GenAI can make errors, creating authentic opportunities for critical analysis. Prompts to GenAI were intentionally simple, mirroring how students typically interact with these tools. Examples include: "det of this 4x4 matrix [matrix data]" or "Tell me the solution of this linear system [system data]".

This authenticity was crucial, using sophisticated prompting techniques would have created a false scenario disconnected from students' actual GenAI usage.

The intervention followed the Safe AI in Education Manifesto principle (Alier Forment et al., 2024). Students did not directly use GenAI with personal accounts; all interactions occurred through the instructor's projected screen. This approach addressed privacy (no student data processed), equity (equal access regardless of resources), transparency (students informed about AI versions used), and critical distance (encouraging collective analysis rather than individual dependency).

Finally, the intervention was designed around four interconnected dimensions:

1. Disciplinary: Improving mathematical understanding through problem-solving and error analysis;
2. Digital: Promoting conscious, critical use of GenAI tools;
3. Andragogical / Pedagogical: Enhancing internal motivation and self-efficacy;
4. Metacognitive: Developing transversal reflection and self-regulation skills.

4. Implementation and Results

The first "discuss the screen" session generated considerable surprise. Many students had assumed GenAI tools were infallible for mathematical calculations. When ChatGPT produced a determinant value different from many students' calculations, the initial reaction was confusion. Rather than immediately reveal-

ing the correct answer, the instructor facilitated discussion: "How did you arrive at your answer? Can you explain each step?" Students began articulating their reasoning, identifying potential sign errors, transposition mistakes, or computational errors.

Importantly, GenAI did not always produce errors. In several cases, the AI's solution was correct, while some students had made mistakes. This proved equally valuable pedagogically: when there was no convergence within the class, collective reasoning helped identify errors made by some students. These instances reinforced correct problem-solving approaches and demonstrated that the methodology served not only to critique AI but to develop mutual learning within the student community.

Following the initial session, the instructor provided a simplified explanation of how Large Language Models function—predicting likely tokens based on statistical patterns rather than "understanding" mathematics. This demystification helped students appreciate both capabilities and fundamental limitations of GenAI. Students began to recognize that GenAI's errors weren't random failures but systematic limitations stemming from its statistical nature.

Subsequent sessions showed evolving sophistication in students' critical analysis. Initially, students identified primarily computational errors, like incorrect arithmetic in determinant expansions, sign errors in matrix operations, or mistakes in Gaussian elimination steps and limit calculations. Over time, students began identifying higher-level conceptual errors. For instance, when GenAI occasionally generated explanations suggesting "a matrix can have two different determinants", students recognized the conceptual impossibility. Similarly, when calculating the area bounded by a curve, GenAI sometimes directly computed the definite integral without checking the sign of the function, thus returning the signed area rather than the actual geometric area; a subtle but important distinction that requires proper setup of the definite integral. This shift from detecting computational mistakes to identifying logical and methodological inconsistencies represents significant cognitive development.

Students progressively developed *dialogical mathematical competence:* the ability to articulate mathematical reasoning clearly and precisely. One student's reflection captured this: "Explaining to someone who doesn't know but gives the impression of knowing is very complex". This observation reveals deep metacognitive awareness. Unlike explaining to an instructor or peer, explaining to GenAI requires assuming nothing and making all reasoning explicit. Throughout the semester, the instructor consistently emphasized this principle of explicit communication, and GenAI sessions provided concrete practice.

The most striking outcome appeared in final exam papers. Compared to pre-

vious years, there was a noticeable overall increase in explanatory comments between mathematical steps. Students wrote annotations like "applying linearity property", "since the determinant is zero, the system has infinite solutions", or "checking the rank condition before proceeding". These comments indicate metacognitive awareness with students monitoring their own reasoning and making implicit knowledge explicit.

Student engagement remained high throughout the semester. The first exam session saw higher attendance than in previous years, suggesting students actively followed the course. Despite increased participation, the pass rate remained consistent with previous first sessions, meaning that, in absolute numbers, more students successfully passed. Notably, this occurred despite the exam being scheduled one week earlier than usual, suggesting students were better prepared and more confident.

Students developed genuine critical thinking skills regarding GenAI outputs. The types of errors students learned to identify included computational errors (arithmetic mistakes, sign errors, transposition errors), procedural errors (incorrect application of algorithms, skipped steps), and conceptual errors (logically impossible statements, misunderstanding of mathematical properties). This progression of critical thinking capabilities aligns with Bloom's taxonomy (Bloom, 1956), moving from remembering and applying procedures to analyzing, evaluating, and creating explanations of mathematical reasoning. The intervention positively impacted motivation and self-efficacy. By demonstrating that they could identify errors that sophisticated AI made, students gained confidence in their own mathematical judgment. The collaborative nature of sessions created a supportive learning environment where errors were treated as learning opportunities rather than failures.

5. Discussion

This exploratory experience demonstrates that GenAI, when thoughtfully integrated, need not undermine mathematical education. Rather than pretending GenAI doesn't exist or prohibiting its use, the intervention embraced its presence while highlighting its limitations. The key pedagogical move was shifting from "AI as solution provider" to "AI as fallible collaborator requiring supervision". These reframing positions human mathematical understanding as essential rather than obsolete.

Skemp's framework proved particularly valuable for analyzing the pedagogical dynamics. GenAI's instrumental competence without relational understand-

ing created perfect teaching scenarios. When ChatGPT correctly applied cofactor expansion but made a sign error, it demonstrated instrumental understanding's insufficiency. When it generated conceptually impossible explanations, it revealed the absence of relational understanding. Students, by contrast, were developing both forms of understanding simultaneously.

The "discuss the screen" methodology exemplifies an innovative application of scaffolding theory. Rather than providing direct instructional support, the approach uses GenAI outputs as cognitive scaffolds that prompt metacognitive engagement. Key features include:

– structured comparison: presenting parallel solutions reduces self-assessment complexity,
– progressive complexity: gradual progression from computational to conceptual errors,
– collaborative support: whole-class discussion provides peer scaffolding,
– fading support: instructor interventions become less directive over time, and
– metacognitive prompting: requirement to explain "why" and "how".

GenAI serves as "provocative scaffolding", that is, it doesn't always provide correct support but rather provokes cognitive engagement through its fallibility.

The methodology presents several challenges. Time management is critical because facilitating discussion cannot be rushed, requiring a careful balance between content coverage and depth of engagement. Instructor improvisation is demanding due to the unpredictability of GenAI outputs, requiring deep content knowledge and pedagogical agility. Assessment alignment requires rethinking, trying to reward metacognitive competencies alongside correct answers. Scalability considerations are important: the approach was designed for 150 students with a variety of backgrounds, and it may require adaptation for much homogeneous classes. This study has significant limitations that must be acknowledged. The lack of systematic assessment means no pre/post measures of metacognitive skills were administered and suggests directions for future research. The absence of a control group makes it impossible to isolate the specific contribution of the "discuss the screen" methodology. Limited longitudinal tracking means long-term impact remains unknown. The observational rather than empirical approach reflects informed professional judgment rather than rigorous data collection. Multiple confounding variables could explain the observations.

6. Conclusion

This educational experience demonstrates that GenAI can be transformed into a powerful pedagogical tool for developing critical thinking and relational understanding. The "discuss the screen" methodology created authentic learning situations where students developed metacognitive awareness, dialogical competence, and genuine mathematical judgment.

Key success factors included treating GenAI outputs as fallible rather than authoritative, creating collaborative environments where errors are learning opportunities, emphasizing the irreplaceable value of human understanding, consistently requiring articulation of mathematical reasoning, maintaining ethical transparency, and leveraging scaffolding principles.

The intervention addressed multiple competency dimensions simultaneously preparing students not only to pass mathematics exams but to function as critical, reflective professionals in an AI-augmented world. As GenAI tools become increasingly sophisticated and ubiquitous, education must evolve beyond simply transmitting knowledge or skills that machines can replicate. Instead, we must cultivate distinctively human capacities: critical judgment, metacognitive awareness, creative problem-solving, and the ability to evaluate and contextualize information from any source.

This exploratory experience in mathematical foundations for architecture students offers one model for that evolution, demonstrating that the challenge of GenAI can indeed become an opportunity for deeper, more meaningful learning. However, substantial empirical research remains necessary to validate these preliminary observations and develop evidence-based practices. The approach is currently being extended during the 2025/26 academic year to other service mathematics courses, working across different learning moments beyond exercise sessions and integrating discipline-specific AI tools to further explore how GenAI can enhance mathematical education across diverse student populations.

Acknowledgement

The author would like to thank Prof. De Gregorio and the students of "Istituzioni di Matematiche" (a.y. 2024/25) at Politecnico di Torino for their active participation and engagement in the revised exercise sessions.

During the preparation of this work, the author used Claude Sonnet 4.5 (Anthropic) to proofread the writing process. After using it, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Alier Forment, M., Garcia Peñalvo, F., Casañ Guerrero, M. J., Pereira, J. A., & Llorens Largo, F. (2024). *Safe AI in Education Manifesto* (Version 0.4.0). Retrieved from https://manifesto.safeaieducation.org

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: David McKay Company.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Lawrence Erlbaum Associates, Inc.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook for Research on Mathematics Teaching and Learning* (pp. 334-370). New York: MacMillan.

Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching*, 77(1), 20-26.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89-100.

II.8
Ripensare la valutazione come processo nel percorso di apprendimento. Spunti per una riflessione di ricerca fenomenologica integrata con l'AI

Rosamaria Nicassio
Phd Candidate, Università degli studi di Bari "Aldo Moro"
r.nicassio4@phd.uniba.it

Abstract

Il presente contributo permette di approfondire e analizzare un modello formativo che integra DUGI e AI nei contesti universitari, con l'obiettivo di proporre un nuovo dispositivo didattico-pedagogico capace di promuovere apprendimento esperienziale, inclusivo, abilitando una valutazione integrata dei processi formativi. La ricerca, che si colloca all'interno del corso di Filosofia della mente (CdL in Scienze pedagogiche), si è proposta di sviluppare un ambiente formativo capace di leggere e valutare la complessità dei processi di apprendimento, valorizzando le dimensioni relazionali: attraverso l'osservazione fenomenologica integrata con l'LLM-based role play, è stato possibile descrivere e interpretare il vissuto degli attori durante l'interazione, rilevare le strutture essenziali dell'esperienza includendo un avatar digitale all'interno del gruppo di confronto.

Parole chiave: Fenomenologia; formazione; valutazione integrata; DUGI; AI.

This contribution analyze a training model that integrates DUGI and AI in university, with the aim of proposing a new educational and pedagogical device which promotes experiential, inclusive learning, enabling an integrated evaluation of training processes. The research, developed in the *Philosophy of Mind* course (Master's Degree in Pedagogical Sciences), aims to develop an educational environment which recognise the complexity of learning processes focusing on their relational dimensions: Through phenomenological observation integrated with LLM-based role play, it was possible to describe and interpret the participants' lived experience during interaction and to identify the essential structures of their experience, also by incorporating a digital avatar within the relation.

Keywords: Phenomenology; education; integrated assessment; DUGI; AI.

1. Fondamenti teorico-metodologici

Nel campo dell'educazione, la valutazione può essere interpretata come un insieme di pratiche finalizzate a esprimere un giudizio circa la distanza che intercorre tra obiettivi dichiarati e condizioni reali, al fine di orientare interventi mirati a ridurre lo scarto identificato (Corsini, 2023). La fase dedicata alla misurazione della distanza tra gli obiettivi perseguiti e la realtà effettiva parte dalla scelta dell'oggetto di osservazione e dall'identificazione delle modalità di osservazione; successivamente, il valutatore decide come interpretare i dati raccolti e in che forma restituirli operando delle scelte assiologiche (*Ibidem*).

Nell'ambito educativo è necessario riconoscere che la pretesa di una valutazione pienamente oggettiva rappresenta un'illusione legata ad almeno due ordini di ragioni: da un lato, ogni processo valutativo implica l'adozione di indicatori per misurare gli apprendimenti e ciò comporta l'inevitabile attribuzione di un giudizio di valore da parte del valutatore, il quale compie scelte interpretative legate alla propria visione del mondo, alle aspettative e agli obiettivi perseguiti; dall'altro lato, anche la decisione su che cosa misurare e come farlo costituisce già un atto valutativo, rendendo la misurazione un momento essenziale della valutazione che nasce con essa, e in essa confluisce (Visalberghi, 1955).

L'impossibilità di una valutazione oggettiva è quindi legata alla esistenza di valori, i quali, se riconosciuti come dimensioni relative, storiche, permettono di interpretare la valutazione come strumento (utile a orientare azioni future, tanto dell'insegnante quanto degli studenti) e come un'assunzione condivisa di responsabilità pedagogica.

Una simile consapevolezza implica la necessità di riconoscere l'esistenza di distorsioni sistematiche nei processi di misurazione e di valutazione. La ricerca docimologica ha evidenziato in modo consistente quanto sia complesso ottenere misure affidabili: uno strumento può essere considerato affidabile solo se, in condizioni identiche, restituisce risultati simili. Tuttavia, l'affidabilità non coincide con la validità, poiché uno strumento, pur coerente nelle sue misurazioni, risulta del tutto inadeguato allo scopo valutativo per cui viene utilizzato: le prove standardizzate, ad esempio, pur essendo costruite con elevati standard tecnici, risultando appropriate su larga scala, non sono progettate per valutare con accuratezza le competenze del singolo studente; impiegarle a tale fine è metodologicamente fuorviante (Corsini, 2023). Nella pratica accademica, dunque, l'inaffidabilità e l'inadeguatezza degli strumenti di misurazione rischiano di tradursi in una significativa frattura nella fiducia degli studenti verso i processi valutativi, con implicazioni educative profonde (*Ibidem*).

In virtù di quanto detto, diventa essenziale riconoscere, all'interno del pro-

cesso valutativo, la presenza di componenti soggettive (Domenici, 1993), le quali producono un apprendimento esperienziale quando, promuovendo partecipazione e trasparenza, realizzano la natura intersoggettiva (Corsini, 2024) della persona permettendo di concepire la valutazione non come mera certificazione della prestazione (Trinchero 2023), ma come un processo decisionale arricchito dal confronto tra i soggetti. L'intersoggettività costituisce, dunque, la condizione generativa per un apprendimento collaborativo (Stahl, 2015): è attraverso la sequenzialità degli enunciati che si realizza un terreno condiviso, all'interno del quale il valore di ogni contributo si situa nel mondo discorsivo co-costruito dal gruppo.

In questo scenario, è possibile ripensare radicalmente i processi valutativi: se l'apprendimento è un fenomeno intrinsecamente intersoggettivo, anche la valutazione deve essere ripensata in chiave gruppale al fine di costruire significati e avanzamenti cognitivi collettivi attraverso strumenti capaci di rilevare le dinamiche interattive.

L'esperienza formativa della Didattica Universitaria Gruppo-Interattiva (DUGI), condotta dal Centro Interuniversitario di Ricerca CIRLaGE all'interno dell'ateneo barese, ha permesso di individuare alcuni elementi strutturali che definiscono i fondamenti teoretici di e per un apprendimento intersoggettivo, valorizzando il contesto gruppale e la rete di scambi comunicativi e simbolici come condizioni che rendono possibile, al contempo, la stabilità e la trasformazione del processo formativo (De Mita & Modugno, 2020).

Un approccio intersoggettivo per la valutazione, che trova nella DUGI la sua applicazione, fonda la propria validità scientifica su presupposti teoretici fenomenologici che permettono di leggere l'esperienza nel superamento dell'atteggiamento solipsistico dell'uomo e nell'incontro tra i soggetti.

La fenomenologia è descritta come lo studio dei fenomeni nel loro modo di presentarsi all'esperienza soggettiva, nel modo in cui vengono percepiti, compresi e investiti di significato all'interno della coscienza (Smith, 2025), nonché come l'indagine dell'esperienza vissuta dell'individuo nel mondo.

Il progetto fenomenologico di Edmund Husserl, avviato nel primo Novecento, mira a riconoscere pari dignità all'esperienza oggettiva e a quella soggettiva; la sua opera, infatti, culmina nell'interesse per una fenomenologia pura, intesa come ricerca di un fondamento universale per la filosofia e per la scienza (Laverty, 2003). Husserl rifiuta l'assolutizzazione positivista dell'osservazione oggettiva della realtà esterna, sostenendo che l'autentico oggetto della ricerca scientifica debba essere il fenomeno così come si presenta alla coscienza dell'individuo: ne deriva che l'indagine fenomenologica non deve essere guidata da assunzioni precostituite (Husserl, 2002). Afferma Husserl: "Ogni conoscenza

autentica e, in particolare, ogni conoscenza scientifica si fonda sull'evidenza interiore" (Husserl, 2015), evidenza (ciò che si manifesta alla coscienza) come luogo primario in cui il fenomeno deve essere indagato; compito del fenomenologo consiste quindi nel "descrivere le cose in se stesse, permettendo a ciò che si presenta di entrare nella coscienza ed essere compreso nei suoi significati ed essenze alla luce dell'intuizione e dell'autoriflessione" (Moustakas, 1994).

In quest'ottica, al ricercatore non è richiesto di porsi come osservatore neutrale o esterno rispetto all'esperienza indagata, ma di assumere un atteggiamento fenomenologico riflessivo che consenta di sospendere le prese di posizione naturali e le interpretazioni preconcette, senza per questo annullare il ruolo della propria evidenza soggettiva. La distanza critica dai propri vissuti non implica infatti una loro esclusione dal processo conoscitivo, ma il loro ricollocamento entro l'orizzonte dell'epoché, che permette di orientare l'attenzione verso le modalità di manifestazione del fenomeno così come esso si dà nelle esperienze dei partecipanti. È in questo movimento riflessivo, che coinvolge tanto la coscienza del ricercatore quanto quella dei soggetti coinvolti, che diventa possibile il passaggio dalle descrizioni fattuali all'individuazione delle essenze, attraverso l'accesso allo stato dell'Io trascendentale (Neubauer et al., 2019).

2. La sperimentazione: contesto, strumenti di ricerca e analisi dei dati

Nell'ambito del corso di Filosofia della mente del CdL in Scienze pedagogiche, l'introduzione delle schede di rilievo fenomenologico rappresenta un'applicazione dell'impianto teorico fin qui delineato.

Le schede sono concepite come strumenti quali-quantitativi finalizzati a rilevare i dati di realtà e a sostenere gli studenti nell'analisi del proprio vissuto cognitivo, emotivo e corporeo durante le attività: attraverso un protocollo strutturato, da un lato esse favoriscono la rilevazione dei dati oggettivi "sulla scena", dall'altro offrono uno spazio di esplicitazione e condivisione che consente al gruppo di articolare significati comuni, riconoscere divergenze e cogliere le trasformazioni generate dall'incontro.

Nella fattispecie, le schede assolvono a una funzione valutativa centrata sull'esperienza vissuta, includendo l'AI come componente dell'ambiente interattivo entro cui si sviluppano i processi di apprendimento. In tale cornice, esse consentono di rilevare il modo in cui gli studenti si collocano fenomenologicamente nell'interazione, favorendo al contempo una postura riflessiva che li rende osservatori attivi dei propri processi esperienziali e offrendo al conduttore elementi utili per monitorare lo sviluppo di competenze riflessive, analitiche e collabora-

tive. In questa prospettiva, la valutazione non riguarda l'efficacia tecnica dell'AI, ma la capacità degli studenti di riflettere sul proprio modo di esperire e significare l'interazione, riconoscendo l'AI come uno dei fattori che contribuiscono a modellarla.

L'attività svolta si è articolata due fasi:

1. Interazione tra attori umani;
2. Interazione tra umano e AI.

Nella prima fase, attraverso la pratica del role playing, l'attore umano 1 è stato chiamato a interagire con l'attore umano 2 nella simulazione di una relazione di cura: l'attore umano 1 ha interpretato una persona (dall'età a scelta) con difficoltà nell'espressione emotiva, chiusa in se stessa, incapace di interagire con gli altri; l'attore umano 2 ha interpretato il ruolo di un pedagogista in grado di intercettare le difficoltà dell'interlocutore, capace di offrire un ascolto non giudicante e accogliente, e con il quale co-creare e generare idee, micro-narrazioni.

La seconda fase, poi, ha apportato una sola grande modificazione: è stata l'AI a interpretare il ruolo del pedagogista dopo aver ricevuto il seguente prompt:

> *"Da questo momento, tu, AI, Interpreti il ruolo di un pedagogista che accompagna una persona in una situazione della vita quotidiana. Il tuo compito è sostenere l'esplorazione di pensieri ed emozioni, aiutare a mettere ordine favorendo l'autonomia decisionale, senza giudicare, dare ordini o sostituirti alle scelte dell'altro. Agisci come facilitatore.*
> *La scena si svolge in una stanza silenziosa. L'altra persona, interpretata da me, fatica a esprimere le proprie emozioni e tende all'isolamento. L'obiettivo del role-play è co-costruire un dialogo in cui io esprimo dubbi e difficoltà e tu mi consenti di esplorare il vissuto attraverso domande aperte, rispettando una gradualità intenzionale. Inizia chiedendomi: «Come stai?»".*

Le schede utilizzate per la raccolta dati son state impiegate per leggere l'esperienza educativa su più livelli (descrittivo, percettivo, emotivo, riflessivo-interpretativo) e son state così articolate:

– Dati di realtà (dati assertivi) del contesto: l'osservatore è chiamato a descrivere in modo il più possibile oggettivo ciò che accade durante l'interazione, soffermandosi sul setting, sulle azioni compiute dai partecipanti;
– Fattori cronotopici, ovvero le dimensioni spaziali e temporali entro cui si sviluppa l'interazione educativa. L'attenzione è rivolta alle posture educative che

emergono nel tempo, alla durata delle diverse fasi dell'interazione, alle trasformazioni dello spazio relazionale (prossemica, posizionamento dei corpi, distanze) e al modo in cui tali elementi incidono sull'andamento della relazione;
- Fattori relativi alla comunicazione, relativi al linguaggio utilizzato, alla struttura dello scambio dialogico, alla gestione dei turni di parola, delle pause e dei silenzi, nonché all'eventuale presenza di distorsioni comunicative;
- Fattori di contagio empatico: vengono rilevati fenomeni di rispecchiamento, amplificazione o attenuazione emotiva, nonché le tonalità affettive che influenzano l'agire educativo e la percezione del problema;
- Fattori di resistenza, elementi che ostacolano o deviano il processo educativo. Questa sezione permette di individuare le dinamiche che "tradiscono" il contesto di riferimento, rendendo visibili le tensioni, le rigidità e le difficoltà che emergono nell'agire educativo;
- Fattori percettivi del corpo in relazione al fenomeno: l'osservatore è chiamato a descrivere la propria esperienza sensoriale e corporea durante l'osservazione. Vengono considerate le sensazioni fisiche, le risonanze emotive e l'impatto che posture, gesti hanno avuto sulla comprensione del fenomeno educativo;
- Processi di significazione: l'osservatore è invitato a rielaborare riflessivamente l'esperienza osservata, attribuendole significato e individuandone le ricadute sulla propria pratica professionale.

Dall'analisi dei dati emerge che la struttura temporale dell'interazione educativa costituisce un elemento centrale del processo formativo. Nelle interazioni umane, il tempo si configura come dimensione costitutiva dell'apprendimento: la possibilità di sostare nei silenzi, rispettare esitazioni e ritmi soggettivi favorisce una progressiva elaborazione del vissuto e l'apertura emotiva. Al contrario, nell'interazione con l'AI il tempo appare statico e la comunicazione, pur formalmente continua, risulta lineare e poco trasformativa, non accompagnando il soggetto in un reale processo di rielaborazione dell'esperienza (Oritsegbemi, 2023).

Analogamente, l'interazione umana attiva processi di rispecchiamento emotivo e coinvolgimento corporeo che consentono alla persona di sentirsi riconosciuta, attraverso segnali verbali, paralinguistici e corporei che costruiscono un clima di fiducia e accoglienza. Nell'interazione con l'AI, invece, le risposte, sebbene talvolta coerenti sul piano contenutistico, non restituiscono le sfumature affettive implicite nella comunicazione, generando percezioni di distanza e freddezza relazionale, osservabili anche nella staticità corporea dell'attore umano. Inoltre, mentre l'interazione umana favorisce la trasformazione del vissuto in

narrazione e la costruzione di significati condivisi, nell'interazione con l'AI tali processi si arrestano a un livello descrittivo o prescrittivo (Nicolella, 2025).

I risultati confermano che la relazione educativa non è riducibile a uno scambio informativo o a un dispositivo di problem solving, ma si configura come un processo complesso che integra dimensioni temporali, corporee, emotive e simboliche, rispetto alle quali l'intelligenza artificiale mostra limiti strutturali. Tuttavia, se collocata entro una cornice metodologica esplicita, l'AI può assumere una funzione complementare e non sostitutiva, operando come supporto alla verbalizzazione e come strumento di organizzazione formale del materiale esperienziale. In coerenza con l'impostazione fenomenologica, che riconosce il significato come esito di un atto intenzionale e intersoggettivo, l'AI non interpreta né comprende l'esperienza, ma può agire come "filtro ordinatore", riducendo la dispersione del materiale verbale senza intervenire direttamente nei processi relazionali ed emotivi, ma preparando il territorio per l'esplorazione (Samuel et al., 2022). La trasformazione simbolica del vissuto resta prerogativa dell'incontro educativo umano, luogo in cui l'esperienza può essere interrogata e rielaborata attraverso la risonanza corporea e intersoggettiva (Chung & Pennebaker, 2011). La relazione educativa propriamente detta inizia, infatti, solo quando il materiale organizzato viene reintrodotto nello spazio dell'incontro umano, dove il vissuto può essere interrogato e trasformato attraverso processi di risonanza corporea incarnata.

3. Conclusioni

Dallo studio condotto emerge la possibilità di una valutazione integrata che incontra fenomenologia e AI, intesa come processo interpretativo che tiene insieme l'esperienza vissuta e i materiali che da essa emergono, senza ridursi a una procedura tecnicistica. L'intelligenza artificiale può assumere una funzione di supporto non invasiva, contribuendo all'organizzazione e alla restituzione strutturata del materiale verbale e narrativo, senza intervenire nei processi interpretativi o nel giudizio educativo che restano atti relazionali e situati; la valutazione propriamente detta resta, infatti, un atto eminentemente relazionale e situato, che implica la capacità di cogliere il senso dell'esperienza nel suo divenire leggendo le trasformazioni soggettive. L'integrazione dell'AI non sostituisce il giudizio professionale, ma ne sostiene l'esercizio, rendendo più leggibile la complessità dei dati e preservando la centralità dell'esperienza e della relazione come fondamenti dell'atto valutativo.

Riferimenti bibliografici

Chung, C., & Pennebaker, J. (2011). The psychological functions of function words. In S. Fiske, D. Gilbert, G. Lindzey (Eds.), *Handbook of social psychology* (pp. 343–359). Hoboken (NJ): Wiley.

Corsini, C. (2023). *La valutazione che educa. Liberare insegnamento e apprendimento dalla tirannia del voto*. Milano: FrancoAngeli.

Corsini, C. (2024). Una valutazione col pilota automatico? Una riflessione sulle cose che possiamo guadagnare e quelle che rischiamo di perdere impiegando l'intelligenza artificiale nei processi valutativi. *Journal of Educational, Cultural and Psychological Studies*, 30, 197–208. https://doi.org/10.7358/ecps-2024-030-corc

De Mita, G., & Modugno, A. (2020). *Insegnare filosofia in università. Riflessioni teoretiche verso nuovi scenari metodologici*. Milano: FrancoAngeli.

Domenici, G. (1993). *Manuale della valutazione scolastica*. Roma-Bari: Laterza.

Husserl, E. (2015). *La crisi delle scienze europee e la fenomenologia trascendentale* (trad. di E. Filippini). Milano: Il Saggiatore.

Husserl, E. (2002). *Idee per una fenomenologia pura e una filosofia fenomenologica. Libro primo: Introduzione generale alla fenomenologia pura* (a cura di V. Costa). Torino: Einaudi.

Laverty, S. M. (2003). Hermeneutic phenomenology and phenomenology: A comparison of historical and methodological considerations. *International Journal of Qualitative Methods*, 2, 1–29.

Moustakas, C. E. (1994). *Phenomenological research methods*. Thousand Oaks (CA): SAGE.

Neubauer, B. E., Witkop, C. T., & Varpio, L. (2019). How phenomenology can help us learn from the experiences of others. *Perspectives on Medical Education*, 8 (2), 90-97.

Nicolella, D. (2025). I sentimenti di una macchina: empatia umana vs. empatia artificiale nell'interazione educativa. *Cultura pedagogica e scenari educativi*, 3 (1), 67–74. https://doi.org/10.7347/spgs-01-2025-08

Oritsegbemi, O. (2023). Human intelligence versus AI: Implications for emotional aspects of human communication. *Journal of Advanced Research in Social Sciences*, 6 (2), 76–85. https://doi.org/10.33422/jarss.v6i2.1005

Samuel, J., Kashyap, R., Samuel, Y., & Pelaez, A. (2022). Adaptive cognitive fit: Artificial intelligence augmented management of information facets and representations. *International Journal of Information Management*, 65, 102505. https://doi.org/10.1016/j.ijinfomgt.2022.102505

Smith, D. W. (n.d.). Phenomenology. In *Stanford Encyclopedia of Philosophy*. Retrieved December 7, 2025, from https://plato.stanford.edu/entries/phenomenology/

Stahl, G. (2015). Conceptualizing the intersubjective group. *International Journal of Computer-Supported Collaborative Learning*, 10 (3), 223–234.

Trinchero, R. (2023). Assessment as learning in università. Costruire le capacità autovalutative degli studenti. *Pedagogia Oggi*, 21 (1), 43–58.

Visalberghi, A. (1955). *Misurazione e valutazione nel processo educativo*. Milano: Edizioni di Comunità.

II PARTE

Area tematica 2. IA e valutazione nella scuola

II.9
Docenti e GenAI nella valutazione: tra entusiasmo e scetticismo. Evidenze dal Questionario Insegnante INVALSI 2024/25

Paolo Barabanti
Ricercatore INVALSI

Abstract

Questo studio analizza percezioni, pratiche e atteggiamenti dei docenti italiani rispetto all'uso della GenAI nella valutazione, utilizzando i dati del Questionario Insegnante INVALSI 2024/25. Emergono cinque profili – entusiasti, prudenti, propensi, cauti, scettici – che riflettono modelli valutativi e di uso del digitale già presenti nella professionalità docente. Le analisi mostrano associazioni significative tra atteggiamenti verso la GenAI, pratiche didattiche, uso delle TIC e le caratteristiche socio-anagrafiche e professionali.

Parole chiave: GenAI; valutazione scolastica; docenti; Rilevazioni integrative INVALSI; TIC.

This study analyzes Italian teachers' perceptions, practices, and attitudes regarding the use of GenAI in assessment, using data from the 2024/25 INVALSI Teacher Questionnaire. Five profiles emerge – enthusiastic, prudent, inclined, cautious, and sceptical – reflecting assessment models and digital-use patterns already present within teachers' professional practice. The analyses reveal significant associations between attitudes toward GenAI, teaching practices, ICT use, and socio-demographic and professional characteristics.

Keywords: Generative AI; educational assessment; teachers; INVALSI supplementary surveys; ICT.

1. Introduzione

L'introduzione dell'Intelligenza Artificiale – e in particolare dell'IA generativa (d'ora in poi GenAI) – sta trasformando in profondità le pratiche di insegnamento e valutazione, aprendo al tempo stesso nuove opportunità e rilevanti criticità. Da un lato, questi strumenti permettono di automatizzare attività onerose, generare rubriche, quesiti o materiali personalizzati e possono potenziare l'*assessment for learning* attraverso *feedback* immediati, osservazioni continue dei processi e compiti più autentici e contestualizzati (Hooda et al., 2022; Swiecki et al., 2022; Puentedura, 2010; 2012). Quando percepita come utile dal facile utilizzo, l'IA tende a essere adottata dagli insegnanti come supporto professionale, in linea con i modelli classici di accettazione tecnologica (Davis, 1989).

Accanto ai potenziali vantaggi, emergono però rischi docimologici legati a errori valutativi, incoerenze negli *output*, *bias* algoritmici e opacità dei modelli, sollevando così interrogativi sulla validità e l'affidabilità dei processi di *assessment* supportati da GenAI (Owan et al., 2023). Studi recenti mostrano, ad esempio, che strumenti come ChatGPT possono produrre giudizi non allineati agli obiettivi cognitivi oppure generare item non validi, richiedendo dunque un controllo esperto e un uso critico da parte del docente (Kanık, 2024; Jauhiainen & Garagorry Guerra, 2025). In questa prospettiva, la GenAI appare non come sostituto del giudizio professionale ma come una forma di strumento operativo capace di affiancare il lavoro valutativo, pur entro confini chiari di responsabilità etica e metodologica (Di Padova et al., 2024).

L'integrazione della GenAI nella valutazione richiede dunque nuove competenze: alfabetizzazione algoritmica, consapevolezza critica dei limiti dei modelli, capacità di governare gli strumenti nella progettazione valutativa (Panciroli & Rivoltella, 2023; Ranieri et al., 202). Comprendere atteggiamenti, pratiche e percezioni dei docenti diventa quindi essenziale per orientare verso un uso della GenAI efficace, responsabile e coerente con i principi di equità e di trasparenza.

2. Metodo

Il presente studio vuole analizzare le rappresentazioni e le pratiche dei docenti italiani riguardo all'uso della GenAI nella valutazione degli apprendimenti, con l'obiettivo di identificare variabili che influenzano apertura o resistenza e di individuare profili di atteggiamento professionale.

L'indagine si basa sui dati raccolti attraverso il Questionario Insegnante delle

Rilevazioni Nazionali INVALSI 2024/25, somministrato ai soli docenti delle classi campione selezionate nell'ambito del disegno delle Prove INVALSI. Pur non trattandosi di un campione probabilistico dei singoli docenti, la struttura campionaria garantisce una copertura ampia e diversificata.

Proposto a 10.522 insegnanti, il questionario è stato compilato completamente in 8.212 casi, pari a un tasso di risposta del 78%. In tabella 1 le caratteristiche dei rispondenti:

		v.a.	%
Genere	Maschio	1.037	13%
	Femmina	7.175	87%
Disciplina insegnata	Italiano	3.356	41%
	Matematica	3.069	37%
	Inglese	1.787	22%
Ordine e grado di scuola	Primaria	4.130	50%
	Secondaria I	490	6%
	Secondaria II	3.592	44%
Anni di servizio	Fino a 10	1.456	18%
	da 11 a 20	1.740	21%
	da 21 a 30	2.439	30%
	Oltre i 30	2.577	31%
TUTTI		8.212	100%

Tab. 1 – Caratteristiche e distribuzione del campione
Fonte: elaborazioni su dati del Questionario Insegnante INVALSI a.s. 2024/25

Il Questionario Insegnante raccoglie informazioni sui docenti in merito a opinioni e atteggiamenti verso le Prove INVALSI, esperienze didattiche, possesso di competenze digitali, sviluppo professionale e motivazione. Per il presente studio sono state analizzate in particolare le sezioni riguardanti:

– l'uso attuale o potenziale della GenAI nelle attività professionali e nella valutazione;
– le pratiche valutative adottate;
– le pratiche didattiche realizzate;
– l'uso delle TIC a supporto dell'attività in classe.

La domanda utilizzata come base di analisi per questo studio è: "*L'uso dell'intelligenza artificiale (ad esempio: ChatGPT, Gemini, Copilot) si sta diffondendo sempre più anche tra gli insegnanti. Le è capitato di utilizzarla in ambito professionale?*". Essa prevedeva tre possibili risposte:

– Sì
– No, ma penso che la userò in futuro
– No, non penso che la userò

A seconda dell'opzione selezionata, veniva eventualmente proposta una domanda aperta di approfondimento, utile a raccogliere informazioni qualitative sulle occasioni in cui la GenAI viene già utilizzata o potrebbe essere utilizzata in futuro. In particolare:

– in caso di risposta "Sì" veniva poi presentata la domanda "Per quali momenti dell'attività scolastica e/o per quali funzioni legate alla Sua professione ha già utilizzato strumenti di intelligenza artificiale?";
– in caso di risposta "No, ma penso che la userò in futuro" veniva poi presentata la domanda "Per quali momenti dell'attività scolastica e/o per quali funzioni legate alla Sua professione potrebbe in futuro utilizzare strumenti di intelligenza artificiale?";
– in caso di risposta "No, non penso che la userò" non veniva proposta alcuna ulteriore domanda di approfondimento.

Le risposte aperte hanno così permesso di raccogliere descrizioni sintetiche ma significative sulle pratiche d'uso, dalle quali è stato possibile cogliere l'eventuale riferimento a pratiche o strumenti relativi alla valutazione degli apprendimenti. Sulla base di queste informazioni è stata costruita una variabile dicotomica (uso per la valutazione: sì/no), integrata poi con le risposte chiuse sull'uso attuale o potenziale della GenAI.

L'incrocio di queste informazioni ha così permesso di identificare cinque profili (Fig. 1):

1. Entusiasti – usano la GenAI anche per la valutazione degli apprendimenti
2. Prudenti – utilizzano la GenAI ma non per la valutazione degli apprendimenti
3. Propensi – mostrano apertura verso la GenAI, pur dichiarando che al momento non la utilizzano, e sono intenzionati a utilizzarla anche per la valutazione degli apprendimenti

4. Cauti – esprimono interesse verso la GenAI, anche se al momento non la utilizzano, ma non intendono utilizzarla anche per la valutazione degli apprendimenti
5. Scettici – non hanno intenzione di avvalersi della GenAI nemmeno in futuro

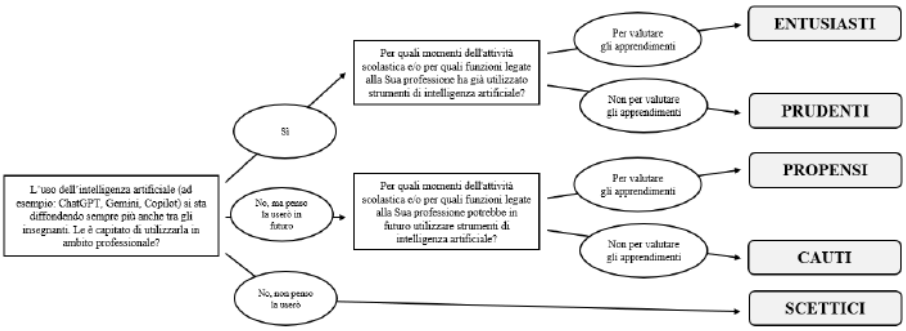


Fig. 1 – Profili creati
Fonte: elaborazioni su dati del Questionario Insegnante INVALSI a.s. 2024/25

Le successive analisi sono state condotte in più fasi. In primo luogo, l'analisi fattoriale ha permesso di identificare cinque dimensioni rilevanti per poter descrivere le pratiche professionali dei docenti:

– un fattore legato alle pratiche didattiche (α = 0,813), relativo ad attività didattiche di natura collaborativa, laboratoriale, multimediale e innovativa;
– tre fattori valutativi:
  • valutazione innovativa (α = 0,708), in caso di pratiche valutative basate su lavori di gruppo e su prodotti multimediali, uso di piattaforme per la valutazione, impiego di compiti di realtà;
  • valutazione costruita (α = 0,747), in caso di utilizzo di prove strutturate e semi-strutturate predisposte dal docente;
  • valutazione tradizionale (α = 0,712), in caso di uso principale di pratiche valutative basate prove scritte (non strutturate) e interrogazioni orali;
– un fattore relativo all'uso delle TIC nella pratica didattica (α = 0,651), nello specifico di piattaforme *e-learning* e di *software.*

Per esaminare le relazioni tra tali profili e le variabili anagrafiche, professionali e didattiche, sono stati applicati il test del χ^2 e il coefficiente V di Cramer; inoltre, sono stati analizzati i residui standardizzati corretti per individuare le celle che

contribuivano in modo significativo alle associazioni osservate. Successivamente, l'analisi MANOVA ha valutato l'effetto multivariato dei profili sulle cinque dimensioni individuate, mentre le ANOVA post-hoc hanno consentito di approfondire le differenze puntuali tra i gruppi.

Il Questionario Insegnante è stato somministrato per mezzo di *LimeSurvey* mentre l'intero processo di analisi è stato condotto tramite il *software SPSS*.

3. Risultati

L'analisi dei dati evidenzia una forte variabilità nelle rappresentazioni, nelle pratiche e nelle intenzioni di uso della GenAI nella valutazione. Fin dalle analisi descrittive (Tab. 2) emerge che circa un terzo circa dei docenti (32%) utilizza già la GenAI, un altro terzo (37%) dichiara di non aver ancora introdotto tali strumenti seppure mostrino interesse verso un possibile impiego futuro mentre il terzo restante (31%) manifesta una posizione di netto rifiuto.

	Sì	No, ma penso la userò in futuro	No, non penso la userò	Totale
Maschio	37%	30%	34%	100%
Femmina	31%	38%	31%	100%
Inglese	39%	35%	26%	100%
Italiano	30%	37%	33%	100%
Matematica	31%	37%	32%	100%
Primaria	24%	41%	35%	100%
Secondaria 1	39%	34%	28%	100%
Secondaria 2	41%	32%	27%	100%
Fino a 10	41%	32%	27%	100%
da 11 a 20	35%	35%	30%	100%
da 21 a 30	32%	40%	28%	100%
Oltre i 30	25%	38%	37%	100%
TUTTI	32%	37%	31%	100%

Tab. 2 – Risposte alla domanda "L'uso dell'intelligenza artificiale si sta diffondendo sempre più anche tra gli insegnanti. Le è capitato di utilizzarla in ambito professionale?".
Fonte: elaborazioni su dati del Questionario Insegnante INVALSI a.s. 2024/25

Nel complesso, la distribuzione dei profili (Tab. 3) mostra una prevalenza di posizioni caute o scettiche. Le differenze risultano marcate per area disciplinare, ordine di scuola e anzianità di servizio mentre per il genere le differenze sono più contenute.

	Entusiasti	Prudenti	Propensi	Cauti	Scettici	Totale
Maschio	14%	20%	4%	31%	30%	100%
Femmina	13%	19%	6%	34%	28%	100%
Inglese	7%	17%	4%	37%	35%	100%
Italiano	18%	20%	6%	28%	28%	100%
Matematica	19%	22%	6%	27%	27%	100%
Primaria	16%	21%	5%	25%	34%	100%
Secondaria 1	12%	19%	5%	33%	31%	100%
Secondaria 2	17%	24%	4%	27%	27%	100%
Fino a 10	9%	16%	4%	33%	37%	100%
da 11 a 20	16%	23%	5%	30%	26%	100%
da 21 a 30	12%	18%	4%	33%	33%	100%
Oltre i 30	12%	19%	5%	32%	32%	100%
TUTTI	13%	19%	5%	32%	31%	100%

Tab. 3 – Profili creati
Fonte: elaborazioni su dati del Questionario Insegnante INVALSI a.s. 2024/25

Per esplorare la relazione tra i profili di atteggiamento verso la GenAI e le caratteristiche dei docenti, sono state condotte diverse analisi tramite test del χ^2 di indipendenza[1]. Il test evidenzia associazioni statisticamente significative per tutte le variabili (Tab. 4), confermando così che la distribuzione dei profili non è casuale ma varia in funzione delle caratteristiche dei docenti.

Per valutare la forza dell'associazione, è stato utilizzato il coefficiente V di Cramer, particolarmente indicato quando si analizzano tabelle di contingenza con più di due categorie. L'entità dell'associazione è compresa tra 0,059 e 0,155, indicando relazioni deboli ma costanti. Tra tutte le variabili, l'ordine di scuola

1 In tutti i casi, le frequenze attese sono risultate adeguate (nessuna cella con frequenze < 5), confermando l'affidabilità dei risultati.

(V = 0,151) mostra le associazioni più consistenti, suggerendo che la collocazione professionale costituisce una dimensione rilevante nel differenziare gli atteggiamenti verso la GenAI.

	χ^2 (df)	p-value	Phi	V di Cramer	Intensità associazione
Materia insegnata	56.342 (8)	< 0,001	0,083	0,059	Debole
Ordine di scuola	374.130 (8)	< 0,001	0,213	0,151	Moderata
Genere	30.657 (4)	< 0,001	0,061	0,061	Debole
Anzianità di servizio	151.758 (12)	< 0,001	0,136	0,078	Debole

Tab. 4 – Risultati del test χ^2 e del coefficiente V di Cramer
Fonte: elaborazioni su dati del Questionario Insegnante INVALSI a.s. 2024/25

Sono stati inoltre analizzati i residui standardizzati corretti (ASR), veri indicatori delle differenze tra frequenze osservate e attese, permettendo così di identificare i gruppi sovra-rappresentati (valori positivi) e sotto-rappresentati (valori negativi) all'interno di ciascuna variabile.

Nella tabella 5, l'intensità dell'ASR è espressa mediante una simbologia sintetica:

– |ASR| I 6 → Molto rilevante (●●●): deviazione molto marcata dalle frequenze attese;
– |ASR| I 4 e < 6 → Rilevante (●●): differenza consistente e statisticamente robusta;
– |ASR| I 2 e < 4 → Moderato (●): differenza significativa ma meno pronunciata;
– |ASR| < 2 → Non significativo: deviazione non statisticamente rilevante (e quindi non indicata in tabella).

L'analisi dei residui standardizzati corretti evidenzia come le differenze più marcate si concentrino principalmente sull'ordine di scuola e sull'anzianità di servizio. In particolare, nella scuola primaria si osserva una forte sovra-rappresentazione dei profili cauti e scettici, a fronte di una netta sotto-rappresentazione degli entusiasti; al contrario, nella secondaria di secondo grado emergono con intensità elevata i profili entusiasti e prudenti mentre risultano significativamente meno presenti quelli cauti e scettici. Anche l'anzianità di servizio mostra un

andamento coerente: i docenti con minore esperienza tendono a collocarsi più frequentemente nei profili entusiasti e prudenti, mentre tra coloro con oltre 30 anni di servizio prevalgono posizioni scettiche. Le differenze per genere e area disciplinare appaiono invece più contenute e circoscritte a singoli profili.

	Entusiasti	Prudenti	Propensi	Cauti	Scettici
Maschio	+3,0 ●	–	–	-5,2 ●●	–
Femmina	-3,0 ●	–	–	+5,2 ●●	–
Inglese	+4,8 ●●	+3,9 ●	–	–	-5,4 ●●
Italiano	-2,4 ●	-2,3 ●	–	–	+2,9 ●
Matematica	–	–	–	–	–
Primaria	-16,1 ●●●	-5,7 ●●	-3,3 ●	+10,1 ●●●	+7,8 ●●●
Secondaria 1	+3,8 ●	–	–	-2,0 ●	–
Secondaria 2	+14,4 ●●●	+5,4 ●●	+2,8 ●	-9,2 ●●●	-7,0 ●●●
Fino a 10	+5,9 ●●	+5,0 ●●	–	-4,4 ●●	-3,7 ●
da 11 a 20	+2,4 ●	–	–	–	–
da 21 a 30	–	–	+2,4 ●	+2,6 ●	-3,7
Oltre i 30	-6,9 ●●●	-4,9 ●●	–	–	+7,7 ●●●

Tab. 5 – Residui standardizzati corretti
Fonte: elaborazioni su dati del Questionario Insegnante INVALSI a.s. 2024/25

Per approfondire come le pratiche didattiche, le modalità valutative e l'uso delle TIC si distribuiscano nei cinque profili individuati, è stata condotta un'analisi MANOVA con cinque variabili dipendenti: pratiche didattiche innovative, valutazione innovativa, valutazione costruita, valutazione tradizionale e uso delle TIC. L'esito multivariato, pur associato a un effetto di entità medio-piccola, indica che i profili di atteggiamento verso la GenAI si associano a differenti configurazioni di pratiche didattiche, valutative e di uso delle tecnologie: Pillai's Trace = 0,058, $F(20, 32824) = 24{,}26$ e $p < 0{,}001$.

A seguito della significatività multivariata, sono state condotte analisi ANOVA per ciascun fattore. Tutti i confronti risultano significativi, confermando che ogni dimensione didattica o valutativa contribuisce alle differenze tra i profili, seppur con intensità variabile (Tab. 6).

Nel loro complesso, le analisi mostrano che:

- i docenti più aperti alla GenAI (Entusiasti e Prudenti) si distinguono per un uso più frequente delle TIC, per l'adozione di pratiche didattiche innovative e per una maggiore propensione verso forme di valutazione autentica e collaborativa;
- i Propensi, pur non utilizzando ancora la GenAI, mostrano profili intermedi, segnati da interesse ma da pratiche meno strutturate;
- i Cauti presentano livelli moderati di innovazione didattica e un uso più contenuto delle TIC, mantenendo un orientamento prudenziale nella valutazione;
- gli Scettici si caratterizzano per un uso significativamente inferiore delle tecnologie, per pratiche didattiche più tradizionali e per una maggiore centralità della valutazione sommativa, suggerendo una postura professionale complessivamente più conservativa.

	p-value	η^2	Effetto	Entusiasti	Prudenti	Propensi	Cauti	Scettici
TIC	< 0,001	0,046	Medio-piccolo	0,252	0,268	-0,008	0,011	-0,280
Pratiche	< 0,001	0,016	Piccolo	0,002	0,206	-0,047	0,031	-0,154
Valutazione Innovativa	< 0,001	0,017	Piccolo	0,043	0,220	-0,069	0,006	-0,150
Valutazione Costruita	0,014	0,002	Molto piccolo	0,058	-0,029	-0,061	0,049	0,001
Valutazione Tradizionale	0,001	0,002	Molto piccolo	0,086	-0,007	0,080	0,051	0,009

Tab. 6 – Risultati ANOVA: differenze tra profili
Fonte: elaborazioni su dati del Questionario Insegnante INVALSI a.s. 2024/25

4. Conclusioni

I risultati di questo studio mostrano come l'atteggiamento dei docenti italiani verso l'uso della GenAI nella valutazione degli apprendimenti sia eterogeneo e stratificato, configurandosi lungo un *continuum* che va dall'entusiasmo allo scetticismo: l'individuazione di cinque profili distinti evidenzia che l'apertura o la resistenza nei confronti della GenAI non può essere interpretata come una semplice reazione a una tecnologia emergente ma riflette orientamenti professionali, didattici e valutativi già consolidati.

I profili empiricamente individuati possono essere letti come l'espressione di diverse modalità di approccio professionale all'innovazione. L'atteggiamento verso la GenAI, in particolare in ambito valutativo, non riguarda esclusivamente l'adozione di un nuovo strumento ma il grado di accettabilità della delega di funzioni valutative a sistemi algoritmici, all'interno di confini percepiti come compatibili con il ruolo e la responsabilità professionale del docente (Confalonieri et al., 2021; Ayanwale et al., 2022).

In linea con i modelli teorici sull'accettazione tecnologica (Davis, 1989) e con la letteratura sull'integrazione critica dell'IA nei contesti educativi (Castaneda & Selwyn, 2018; Selwyn, 2019), lo studio evidenzia associazioni significative tra atteggiamenti più favorevoli verso la GenAI e contesti professionali caratterizzati da pratiche didattiche innovative, uso frequente delle TIC e orientamento verso forme di valutazione autentica e formativa. Viceversa, atteggiamenti di cautela o di rifiuto risultano più frequentemente associati a contesti in cui prevalgono modelli di insegnamento e di valutazione più tradizionali.

Le posizioni di cautela o di scetticismo non appaiono riconducibili a una generica resistenza al cambiamento tecnologico ma si allineano alle criticità epistemologiche ed etiche ampiamente discusse in letteratura, in particolare in relazione all'opacità dei modelli, all'affidabilità degli *output* e al rischio di perdita di controllo professionale sul processo valutativo (Ribeiro et al., 2016; Roncaglia, 2023).

Le analisi multivariate rafforzano questa lettura, mostrando come gli atteggiamenti verso la GenAI si associno in modo coerente a configurazioni già strutturate di pratiche didattiche, valutative e di uso delle tecnologie, piuttosto che a preferenze isolate o contingenti. In particolare, i profili più aperti alla GenAI si collocano in ambienti professionali caratterizzati da un uso più intenso delle TIC e da pratiche valutative orientate alla dimensione formativa e autentica, mentre i profili più cauti o scettici risultano maggiormente legati a modelli didattici e valutativi di tipo tradizionale.

Da questi risultati emergono rilevanti implicazioni per la formazione iniziale e in servizio dei docenti. L'integrazione della GenAI nella valutazione non può essere affidata a interventi formativi meramente tecnici o strumentali ma richiede percorsi mirati capaci di intervenire sulle rappresentazioni professionali, sulle culture valutative e sulle condizioni di esercizio del giudizio docente. La formazione dovrebbe sostenere lo sviluppo di competenze critiche legate all'uso consapevole della GenAI (Cuomo et al., 2022) come strumento di supporto al giudizio professionale, rafforzando la capacità di progettare pratiche valutative coerenti, interpretare criticamente gli output algoritmici e mantenere il controllo epistemico ed etico dei processi di *assessment*.

I risultati discussi delineano un quadro interpretativo articolato ma devono essere letti tenendo conto di alcuni vincoli metodologici che ne delimitano la portata e suggeriscono ulteriori direzioni di approfondimento. Accanto ai contributi offerti, lo studio presenta infatti alcuni limiti. In particolare, il campione, pur ampio e articolato, non è probabilistico e non consente pertanto di generalizzare pienamente i risultati all'intera popolazione docente; inoltre, il ricorso a dati auto-riferiti non permette di verificare direttamente le pratiche effettivamente messe in atto in classe. Infine, la natura dell'indagine non consente di cogliere l'evoluzione degli atteggiamenti nel tempo né di osservare eventuali cambiamenti legati a processi di familiarizzazione o di formazione.

Sviluppi futuri potrebbero includere studi longitudinali, osservazioni in contesto, analisi di artefatti valutativi e disegni di ricerca-azione, al fine di esplorare in modo più approfondito l'impatto concreto dell'integrazione della GenAI nei processi valutativi e nelle pratiche professionali dei docenti.

Riferimenti bibliografici

Ayanwale M.A., Sanusi I.T., Adelana O.P., Aruleba K.D., & Oyelere S.S. (2022). Teachers' readiness and intention to teach artificial intelligence in schools. *Computers and Education: Artificial Intelligence*, 3, Article 100099.

Castaneda L., & Selwyn N. (2018). More than tools? Making sense of the ongoing digitizations of higher education. *International Journal of Educational Technology in Higher Education*, 15(22), 1-10.

Confalonieri R., Penaloza R., Potyka N., & Schlobach, S. (2021). A well founded taxonomy for explainability in artificial intelligence. *Artificial Intelligence Review*, 54(5), 3813-3814.

Cuomo S., Biagini G., & Ranieri M. (2022) Artificial Intelligence Literacy, che cos'è e come promuoverla. Dall'analisi della letteratura ad una proposta di Framework. *Media Education*, 13(2), 161-172.

Davis F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340.

Di Padova M., Tinterri A., Basta A., Amatori G., & Dipace A. (2024). Potenziare il giudizio descrittivo nella scuola primaria con l'uso dell'IA generativa. *IUL Research*, *5*(9).

Hooda M., Rana C., Dahiya O., Rizwan A., & Hossain M.S. (2022). Artificial intelligence for assessment and feedback to enhance student success in higher education. *Mathematical Problems in Engineering*, *2022*.

Jauhiainen J.S., & Garagorry Guerra A. (2025). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, *62*(4), 1377-1394.

Kanık M. (2024). The use of ChatGPT in assessment. *International Journal of Assessment Tools in Education*, *11*(3), 608-621.
Owan V.J., Abang K.B., Idika D.O., Etta E.O., & Bassey B.A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *EURASIA Journal of Mathematics, Science and Technology Education*, *19*(8).
Panciroli C., & Rivoltella P.C. (2023), *Pedagogia algoritmica. Per una riflessione educativa sull'Intelligenza Artificiale*. Brescia: Scholé.
Puentedura R. (2010). *SAMR and TPCK: Intro to advanced practice*. Link: http://hippasus.com/resources/sweden2010/SAMR_TPCK_IntroToAdvancedPractice.pdf.
Puentedura R. (2012). *The SAMR model: Six exemplars*. Link: http://www.hippasus.com.
Ranieri M., Cuomo S., & Biagini G. (2025), *Scuola e intelligenza artificiale. Percorsi di alfabetizzazione critica*. Roma: Carocci.
Ribeiro M.T., Singh S., & Guestrin C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016.
Roncaglia G. (2023). *L'architetto e l'oracolo: forme digitali del sapere da Wikipedia a ChatGPT*. Roma-Bari: Laterza.
Selwyn N. (2019). *Should Robots Replace Teachers? AI and the Future of Education*. Cambridge: Polity Press.
Swiecki Z., Khosravi H., Chen G., Martinez-Maldonado R., Lodge J., Milligan S., Selwyn N., & Gaševi D. (2022). Assessment in the age of Artificial Intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075.

II.10
Conceptual Domains of AI in Education for Neuroinclusive Measurement Using Topic Modelling

Irene Gianeselli
PhD, Research Assistant (Psychometrics), Free University of Bozen-Bolzano, irene.gianeselli@unibz.it
Dario Ianes
PhD, Full Professor (Special Education), Free University of Bozen-Bolzano, dario.ianes@unibz.it
Rita Cosoli
PhD Candidate, University of Geneva, rita.cosoli@unige.ch
Marco Cadavero
PhD Candidate, University of Bologna, marco.cadavero@unibo.it
Gaetano Scianatico
PhD Candidate, University of Bari, gaetano.scianatico@uniba.it
Alessandro Oronzo Caffò
PhD, Associate Professor (Psychometrics), University of Bari, alessandro.caffo@uniba.it,
Andrea Bosco
PhD, Full Professor (Psychometrics), University of Bari, andrea.bosco@uniba.it[1]

Abstract

The increasing integration of artificial intelligence (AI) in educational contexts calls for measurement instruments that adequately capture learners' experiences from a neuroinclusive perspective. This study examines the conceptual structure of recent AI-in-education research to inform early phases of scale development. A systematic Scopus search identified 411 peer-reviewed journal articles published between 2020 and 2025, which were analysed using Latent Dirichlet Allocation topic modelling. Four latent domains were identified, reflecting general educational discourse on AI, analytical research perspectives, measurement and assessment, and medical or clinical training contexts. Lexical frequency analyses further indicated a limited representation of emotional and epistemic constructs.

1 Irene Gianeselli contributed to conceptualization, methodology, formal analysis, and writing – original draft. Dario Ianes contributed to conceptualization. Rita Cosoli, Marco Cadavero, and Gaetano Scianatico contributed to conceptualization and writing – review & editing. Alessandro Oronzo Caffò contributed to methodology and writing – review & editing. Andrea Bosco contributed to methodology, writing – review & editing, and supervision.

Keywords: Artificial intelligence; neurodiversity; topic modelling; scale development; educational psychology.

L'applicazione sempre più diffusa dell'intelligenza artificiale (IA) nei contesti educativi rende necessario disporre di strumenti di misurazione capaci di cogliere in modo preciso l'interazione degli studenti con tali strumenti in una prospettiva neuroinclusiva. Questo studio analizza la struttura concettuale della recente letteratura sull'IA in ambito educativo, con l'obiettivo di orientare le fasi iniziali dello sviluppo di una scala psicometrica. Attraverso una ricerca sistematica su Scopus sono stati individuati 411 articoli scientifici peer-reviewed pubblicati tra il 2020 e il 2025, analizzati mediante tecniche di topic modelling basate su Latent Dirichlet Allocation. Sono emersi quattro domini latenti, riferibili al discorso educativo generale sull'IA, alle prospettive analitiche della ricerca, alla misurazione e valutazione e ai contesti di formazione medica o clinica. L'analisi delle frequenze lessicali ha mostrato una presenza limitata di costrutti emotivi ed epistemici.

Parole chiave: intelligenza artificiale; neurodiversità; topic modelling; sviluppo di scala; psicologia dell'educazione.

1. Introduction

The expanding presence of Artificial Intelligence (AI) in education is reshaping how learners access information, receive feedback, and regulate cognitive processes. Advances such as generative AI and intelligent tutoring systems have shifted AI from a peripheral technological aid to an interactive cognitive agent influencing learning behaviours and decision-making (Bosco, 2024). As AI becomes increasingly embedded in instructional environments, understanding learners' cognitive, emotional, and behavioural experiences has become essential. Existing measurement instruments capture only part of this complexity. Validated scales, including the AI Attitude Scale (AIAS-4; Grassini, 2023), the General Attitudes toward Artificial Intelligence Scale (GAAIS; Schepman & Rodway, 2023), and the Threats of Artificial Intelligence Scale (TAI; Kieslich et al., 2021), primarily assess attitudes or perceived risks. However, they were not developed for educational contexts and do not capture the interplay among cognitive, emotional, and behavioural processes involved in learning with AI. Although research shows that AI can influence metacognition and socioemotional engagement (Pacheco et al., 2025), emotional responses, as well as epistemic and affective–motivational states such as curiosity (Noordewier &

Gocłowska, 2024) and trust (McAllister, 1995), remain insufficiently measured. These gaps are particularly salient in neuroinclusive education (Dark, 2024). Learners who present distinct neurological profiles, including those associated with Autism Spectrum Disorders (ASD), Attention Deficit/Hyperactivity Disorder (ADHD), and dyslexia, as well as cognitive variations such as high intellectual ability, may engage with AI differently due to variation in cognitive functioning (e.g., executive processing, attentional styles, sensory sensitivities, and metacognitive ability). A neurodiversity perspective suggests that AI tools may reduce cognitive load for some profiles while increasing uncertainty or emotional discomfort for others (Hutson, 2024). Existing instruments do not account for neurological variability, which limits their suitability for neuro-inclusive educational research. A further limitation concerns emotional processes. Although AI has been associated with curiosity, motivation, exploratory behaviour, anxiety, and error-related concerns (Montag & Elhai, 2025; Naiseh et al., 2025), these experiences are rarely operationalised in existing measures, constraining investigations of how affect interacts with learning outcomes and self-regulation. This study addresses these gaps by identifying the conceptual domains needed to assess the multidimensional impact of AI in neuroinclusive educational contexts. Through a systematic literature search and computational extraction of semantic patterns, it provides an evidence-informed foundation for developing a new measurement instrument and supports subsequent phases of scale development.

2. Methods

2.1 Literature Search and Corpus Preparation

The Scopus search strategy was operationalised to retrieve peer-reviewed publications simultaneously addressing (a) artificial intelligence, (b) educational contexts, and (c) measurement or psychometric concepts. The search was conducted on titles, abstracts, and keywords (TITLE-ABS-KEY), combining AI-related terms (e.g., *artificial intelligence*, *AI*), educational terms (e.g., *learning*, *student*, *teacher*), and measurement-related terms (e.g., *scale*, *instrument*, *validation*, *factor analysis*). Specifically, records were included if their combined title, abstract, and keyword fields simultaneously contained terms related to (a) psychometric measurement, (b) artificial intelligence, and (c) educational contexts, as operationalised in the Scopus query reported in Appendix A. Automatic filtering was applied sequentially by publication year, document type, language,

and conceptual relevance inferred from text content. No manual screening was performed, ensuring full reproducibility of the document retrieval process. The search was restricted to journal articles published between 2020 and 2025 and written in English. This time frame was selected to capture the period in which AI systems (particularly generative and adaptive AI) became widely accessible in educational settings, leading to a substantive shift in both pedagogical practices and research on learner-AI interaction. Scopus was selected as the sole data source because its journal coverage is substantially broader than that of Web of Science, particularly in educational and technology-enhanced learning research (Mongeon & Paul-Hus, 2016), making it well suited to capturing recent developments in AI-mediated education. Records focusing exclusively on technical AI engineering without educational or psychological relevance were excluded. Titles, abstracts, author keywords, and index keywords of all eligible publications were extracted and concatenated into a unified text corpus. Conceptual inclusion criteria were implemented exclusively through automated text-based filtering rather than manual screening. After filtering, the final corpus consisted of 411 publications, which were retained for text preprocessing and topic modelling.

2.2 Text Preprocessing and Topic Modelling

All text-mining procedures were implemented in R (version 4.5.1) using the *tidyverse*, *tidytext*, *slam*, and *topicmodels* packages. Text preprocessing followed standard natural language processing procedures, including lowercasing, removal of punctuation and stopwords (using the standard *tidytext* stop-word list), tokenisation into unigrams, removal of non-alphabetic tokens, and exclusion of tokens shorter than four characters. Tokens were aggregated into a document–term matrix (DTM) using raw term counts as cell values. The resulting DTM exhibited high sparsity (98.45%), a pattern typical of scientific text corpora and consistent with the assumptions of probabilistic topic modelling. Latent Dirichlet Allocation (LDA) was applied to identify latent semantic structures within the corpus. LDA is a hierarchical Bayesian model in which each document is represented as a mixture of latent topics and each topic as a probability distribution over words (Blei, Ng, & Jordan, 2003). Prior research has shown that LDA can recover coherent conceptual structures in scientific corpora and educational research domains (Griffiths & Steyvers, 2004; Chen et al., 2021). Although structural topic models allow the inclusion of document-level metadata (Roberts, Stewart & Tingley, 2019), classical LDA was considered appropriate

given the exploratory aim of identifying conceptual domains and the absence of theoretically relevant covariates. The number of topics was determined through combined inspection of perplexity trends and qualitative semantic coherence. Alternative solutions were examined. The k = 3 solution collapsed educational, analytical, and measurement-related content into a single broad topic, whereas the k = 5 solution produced semantically overlapping domains with reduced interpretability. In contrast, the four-topic model provided clearer conceptual separation while maintaining parsimony.

2.3 Construct Identification and Model Adequacy Checks

Construct identification involved interpretive examination of topic–word distributions. High-probability terms associated with each latent topic were inspected to identify semantically coherent patterns. Conceptual domains were derived solely from these semantic regularities and the terminology contained in the retrieved literature; no theoretical structures external to the dataset were imposed. Based on the identified domains, a preliminary set of item ideas was drafted to support later validation stages; items are not reported here as they remain in an early exploratory form. Model adequacy, stability, and interpretability were evaluated through multiple diagnostic procedures, drawing on recommendations from probabilistic topic modelling (Blei et al., 2003; Griffiths & Steyvers, 2004), educational applications (Chen et al., 2021), and text-mining quality assurance in systematic reviews (O'Mara-Eves et al., 2015). First, sparsity of the document-term matrix (DTM) was examined. The corpus exhibited high sparsity (0.98), a pattern typical of natural-language educational corpora and consistent with LDA's generative assumptions. Second, model fit across alternative values of *k* was assessed using perplexity. Inspection of the perplexity curve indicated diminishing improvements beyond the selected four-topic solution, supporting its adequacy relative to more complex models. Third, semantic coherence was calculated to evaluate the co-occurrence reliability of high-probability terms using log co-occurrence scores. For the retained four-topic solution, topic-level coherence values ranged from –70.10 to –33.64, reflecting substantial lexical sparsity and thematic dispersion typical of large educational corpora. Log co-occurrence-based coherence metrics produce negative values because they rely on the logarithm of word co-occurrence probabilities. In sparse natural-language corpora, such probabilities are necessarily smaller than 1, resulting in negative log values. Fourth, topic stability was examined through nonparametric bootstrap resampling (50 replications). Stability was evaluated by inspecting the

consistency of high-probability terms across bootstrap solutions, which showed substantial overlap in the most salient terms within each topic. This procedure was intended as a qualitative robustness check rather than a formal inferential test. Finally, interpretability of document–topic distributions was assessed by examining posterior values, which indicated clear dominant-topic probabilities (γ = .39–1.00) and relatively balanced topic prevalence across the four topics.

3. Results

The LDA model identified four latent topics with distinct semantic profiles, as indicated by topic–term probability distributions. One topic primarily captured general educational discourse around AI, characterised by terms such as education, learning, teachers, students, artificial, and intelligence. A second topic reflected research and analytical perspectives, with prominent terms including *analysis*, *research*, *factor*, and *study*. A third topic centred on measurement and assessment, featuring terms such as *scale*, *factor*, *chatgpt*, and *students*. Finally, a smaller yet distinct topic was characterised by medical and clinical terminology, including *medical*, *health*, *clinical*, *study*, and *training*. Document-topic posterior distributions (γ) indicated that Topics 1, 2, and 3 accounted for most documents, whereas Topic 4 represented a smaller cluster. Each document showed a dominant topic assignment (γ range = .39–1.00), supporting the interpretability of the four-topic solution and a clear dominant-topic structure across documents. Analysis of lexical frequencies revealed a strong concentration of terms related to educational, analytical, and AI-related content. The most frequent tokens in the corpus were *education*, *analysis*, *artificial*, *intelligence*, *students*, and *learning*, each occurring over one thousand times in the pre-processed text. In contrast, emotionally relevant terms occurred at substantially lower frequencies. Tokens such as *anxiety*, *engagement*, *emotional*, *trust*, *confidence*, and *fear* together accounted for less than 0.2% of all tokens, as confirmed by aggregate frequency counts, indicating that emotional terminology constituted a marginal proportion of the lexical content in the analysed corpus. Emotional terminology was operationalised as a predefined set of affect-related tokens identified *a priori*. Relative frequencies were calculated as the proportion of total post-preprocessing tokens in the corpus, based on raw term counts rather than weighted measures.

4. Discussion

This study mapped the conceptual structure of recent literature on artificial intelligence in educational contexts with the aim of informing the early phases of developing a neuroinclusive measurement instrument. The topic-modelling results delineate a field that is currently structured around a limited set of dominant conceptual domains, with a clear emphasis on cognitive, analytical, and instructional aspects of AI use. These findings are consistent with accounts describing AI tools as cognitive partners that influence information processing, feedback interpretation, and self-regulatory demands in learning environments (Bosco, 2024; Pacheco et al., 2025). As such, the domains identified here provide an empirically grounded starting point for defining candidate constructs to be operationalised in subsequent scale-development phases. At the same time, the analysis revealed a marked underrepresentation of emotional terminology across the corpus. Although affective responses such as anxiety, trust, curiosity, and confidence are frequently discussed in conceptual and empirical work on human–AI interaction (Montag & Elhai, 2025; Naiseh et al., 2025), these constructs did not emerge as a coherent latent domain in the topic model and accounted for only a minimal proportion of lexical content. This pattern suggests that emotional processes are not systematically foregrounded within the Scopus-indexed literature captured by the present search strategy, despite their documented relevance for engagement, persistence, and help-seeking. From a measurement perspective, this gap highlights the need for future instruments to explicitly incorporate affective dimensions rather than relying solely on cognitively oriented constructs. A related observation concerns the absence of topics explicitly reflecting the involvement of different neurotypes in learning. Despite growing attention to neurodiversity in educational research (Dark, 2024), the corpus did not bear topics centred on neurodivergent learners or neurocognitive variability. This absence is more plausibly attributable to the current empirical focus of *AI in education* research rather than to a lack of theoretical relevance. Given established differences in executive functioning, attentional regulation, and uncertainty processing across neural profiles (Hutson, 2024), the exclusion of such perspectives may constrain understanding of how learners differentially experience and interpret AI-supported learning environments. Addressing this limitation represents an important direction for future research and measurement development. The topic structure also included a domain characterised by medical, clinical, and professional training terminology. This topic is best interpreted considering the disciplinary composition of the retrieved literature and the broad biomedical coverage of Scopus, rather than as a core educational

construct. Similar domain-dependent effects of corpus composition on text-mining outputs have been noted in large-scale methodological studies (O'Mara-Eves et al., 2015). While this domain is peripheral to the primary focus on educational and psychological measurement, its presence underscores the importance of contextualising topic-modelling outputs within the characteristics of the source database and exercising caution when generalising conceptual structures beyond the retrieved corpus. Findings portray a research landscape in which cognitive and analytical dimensions of AI use are well represented, whereas emotional, neurodiversity-relevant, and practice-oriented educational constructs remain comparatively underdeveloped. By systematically identifying these imbalances, the present study contributes an evidence-informed basis for subsequent stages of scale development. In particular, the results support a measurement agenda that moves beyond predominantly cognitive framings to incorporate affective processes and neuroinclusive considerations, thereby better capturing the multidimensional impact of AI in educational contexts. In the subsequent phase of scale development, the empirically derived domains will be translated into theoretically bounded constructs, which will guide item generation through a combined inductive–deductive procedure. Item formulation will be refined through expert content validation prior to formal psychometric testing.

5. Conclusion

The present study offers an evidence-based overview of the conceptual domains shaping current research on AI in education, providing a starting point for developing a novel neuroinclusive measurement instrument. The results highlight areas of conceptual concentration along with gaps (particularly in emotional, confidence-related, neurodiversity-relevant, and practice-oriented educational dimensions) that will require integration during item generation. The four identified domains will serve as an initial scaffold for constructing the exploratory item pool. The next steps include refining items, conducting expert content validation, and examining dimensionality, reliability, and measurement invariance to ensure that the emerging tool can adequately capture the multifaceted impact of AI on learning across diverse learner profiles.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research,* 3, 993-1022.

Bosco, A. (2024, March 15). *IA una questione di metodo?* In *Approcci collaborativi e Intelligenza Artificiale nelle organizzazioni: Il contributo delle Neuroscienze*. IrcCAN – International Research Center for Cognitive Applied Neuroscience, Università Cattolica del Sacro Cuore, Milan.

Chen, X., Zou, D., Xie, H., & Su, F. (2021). Twenty-five years of computer-assisted language learning: A topic modeling analysis. *Language Learning & Technology, 25*(3), 151-185.

Dark, J. (2024). Eight principles of neuro-inclusion; an autistic perspective on innovating inclusive research methods. *Frontiers in psychology*, 15, 1326536. https://doi.org/10.3389/fpsyg.2024.1326536

Grassini S. (2023). Development and validation of the AI attitude scale (AIAS-4): a brief measure of general attitude toward artificial intelligence. *Frontiers in psychology*, 14, 1191628. https://doi.org/10.3389/fpsyg.2023.1191628

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101 Suppl 1*(Suppl 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Hutson, P. (2024). Embracing the irreplaceable: The role of neurodiversity in cultivating human-AI symbiosis in education. *International Journal of Emerging and Disruptive Innovation in Education: VISIONARIUM, 2*(1), 5. https://doi.org/10.62608/2831-3550.1020

Kieslich, K., Lünich, M., & Marcinkowski, F. (2021). The Threats of Artificial Intelligence Scale (TAI): Development, measurement and test over three application domains. *International Journal of Social Robotics, 13*(7), 1563-1577. https://doi.org/10.1007/s12369-020-00734-w

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal, 38*(1), 24-59. https://doi.org/10.2307/256727

Mongeon, P., Paul-Hus, A. The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics* 106, 213-228 (2016). https://doi.org/-10.1007/s11192-015-1765-5

Montag, C., & Elhai, J. D. (2025). The darker side of positive AI attitudes: Investigating associations with (problematic) social media use. *Addictive behaviors reports*, 22, 100613. https://doi.org/10.1016/j.abrep.2025.100613

Naiseh, M., Babiker, A., Al-Shakhsi, S. *et al.* (2025). Attitudes Towards AI: The Interplay of Self-Efficacy, Well-Being, and Competency. *J. technol. behav. sci.* https://doi.org/10.1007/s41347-025-00486-2

Noordewier, M. K., & Gocłowska, M. A. (2024). Shared and unique features of epistemic emotions: Awe, surprise, curiosity, interest, confusion, and boredom. *Emotion, 24*(4), 1029-1048. https://doi.org/10.1037/emo0001314

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, *4*(1), 5. https://doi.org/10.1186/2046-4053-4-5

Pacheco, A. J., Boude Figueredo, O. R., Chiappe, A., & Fontán de Bedout, L. (2025). AI-powered learning analytics for metacognitive and socio-emotional development: A systematic review. *Frontiers in Education,* 10, 1672901. https://doi.org/-10.3389/feduc.2025.1672901

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2), 1-40. https://doi.org/10.186-37/jss.v091.i02

Schepman, A., & Rodway, P. (2023). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human–Computer Interaction*, *39*(13), 2724-2741. https://doi.org/10.1080/10447318.2022.2085400

Appendix A. Operational Scopus Search Strategy

The Scopus search strategy was operationalised to identify peer-reviewed journal articles addressing the intersection of artificial intelligence, educational contexts, and measurement or psychometric constructs. The search was conducted using the *TITLE-ABS-KEY* field to ensure coverage of titles, abstracts, and both author and index keywords. The operational search string was:

```
TITLE-ABS-KEY(
  scale OR questionnaire OR instrument OR psychometric OR validation
  OR “item development” OR “factor analysis” OR CFA OR EFA OR measure
)
AND
TITLE-ABS-KEY(
  “artificial intelligence” OR AI OR “intelligent system”
)
AND
TITLE-ABS-KEY(
  education OR learning OR student OR teacher OR school OR “higher education”
)
AND (PUBYEAR > 2019 AND PUBYEAR < 2026)
AND (LIMIT-TO(LANGUAGE, “English”))
AND (LIMIT-TO(DOCTYPE, “ar”))
```

II.11
Le percezioni degli insegnanti circa l'uso dell'intelligenza artificiale nella valutazione della produzione scritta degli studenti: uno studio esplorativo

Massimo Marcuccio
Professore ordinario di Pedagogia Sperimentale (SSD: PAED – 02/B)
Alma Mater Studiorum – Università di Bologna.
massimo.marcuccio@unibo.it
Maria Elena Tassinari
Assegnista di ricerca in Pedagogia Sperimentale (SSD: PAED – 02/B)
Alma Mater Studiorum – Università di Bologna.
mariaelena.tassinar3@unibo.it

Abstract

L'impiego dell'Intelligenza Artificiale (IA) nei contesti educativi è in rapida espansione, in particolare nelle pratiche di progettazione e valutazione, suscitando al contempo interesse e preoccupazioni tra i docenti. Questo studio esplorativo analizza le percezioni di due insegnanti di scuola secondaria di primo grado rispetto all'uso di ChatGPT nella valutazione della produzione scritta, mettendo a confronto i giudizi professionali con quelli generati dal chatbot. L'analisi delle interviste qualitative evidenzia opportunità, criticità e tensioni che emergono dall'incontro tra pratiche valutative consolidate e automazione intelligente, offrendo indicazioni utili per orientare future riflessioni e ricerche sul ruolo dell'IA nei processi valutativi scolastici.

Parole chiave: Intelligenza artificiale; valutazione degli apprendimenti; insegnanti; ChatGPT.

The use of Artificial Intelligence (AI) in educational contexts is rapidly expanding, particularly in instructional design and assessment practices, and is generating both interest and concern among teachers. This exploratory study examines the perceptions of two lower secondary school teachers regarding the use of ChatGPT in evaluating students' written work, comparing their professional judgments with those produced by the chatbot. The analysis of qualitative interviews highlights the opportunities, challenges, and tensions that arise when established

assessment practices encounter intelligent automation, offering insights to inform future reflections and research on the role of AI in school-based assessment processes.

Keywords: Artificial intelligence; learning assessment; teachers; ChatGPT.

1. Introduzione

L'intelligenza artificiale (IA) sta trasformando in modo profondo e trasversale numerosi ambiti della vita sociale, includendo sempre più stabilmente il settore educativo. La scuola diventa così un laboratorio privilegiato di sperimentazione, nel quale l'IA viene impiegata per potenziare l'apprendimento, personalizzare i percorsi formativi e automatizzare processi chiave quali la progettazione e la valutazione (Goralski et al., 2020). In quest'ultimo ambito, l'introduzione di strumenti basati su IA, soprattutto quelli che si avvalgono dell'IA generativa (IAGen), solleva tuttavia questioni teoriche e metodologiche rilevanti, che impongono un'analisi accurata delle percezioni, delle pratiche e delle dinamiche di interazione tra insegnanti e tecnologie emergenti. La letteratura sulle tecnologie educative evidenzia come l'adozione di nuovi strumenti digitali sia influenzata da un intreccio di fattori cognitivi, sociali e istituzionali. Il *Technology Acceptance Model* (TAM) di Davis (1989) e la *Unified Theory of Acceptance and Use of Technology* (UTAUT) di Venkatesh e colleghi (2003) forniscono utili cornici interpretative per comprendere le condizioni che favoriscono o ostacolano l'integrazione di sistemi automatizzati nella valutazione. A tali prospettive si affianca il contributo di Ertmer (1999), che distingue tra barriere di primo ordine, legate alle risorse infrastrutturali e alla formazione tecnica e barriere di secondo ordine, riconducibili alle convinzioni pedagogiche e alle resistenze epistemologiche degli insegnanti. Nell'ambito dell'*Artificial Intelligence in Education* (AIEd) (Luckin, 2018), queste dimensioni concorrono a delineare un continuum di atteggiamenti nei confronti dell'IA valutativa, che spazia dall'integrazione stabile dell'innovazione al suo uso condizionato, fino al rifiuto o alla limitazione del suo impiego.

All'interno di tale quadro si colloca l'interesse per la valutazione della produzione scritta, ambito che ha conosciuto una profonda evoluzione teorica. Come evidenziato da Huot (2002) e da Huot e O'Neill (2009), fino agli anni Settanta la valutazione della scrittura era inscritta nella tradizione della misurazione educativa, centrata su standardizzazione, affidabilità e validità, assumendo la scrittura come fenomeno quantificabile. A partire dagli anni Settanta, gli studi

sulla composizione e sulla retorica hanno invece messo in luce la natura sociale, situata e processuale della scrittura, criticando la riduzione a criteri oggettivi. Tra gli anni Ottanta e Novanta si consolida l'idea di una validità dinamica e contestuale e si diffondono pratiche valutative quali saggi, portfolio e compiti autentici, con attenzione alla validità costruttiva e consequenziale e al ruolo attivo dell'insegnante nel progettare pratiche coerenti con l'apprendimento. L'avvento delle tecnologie digitali e, più recentemente, dei modelli di IA ha introdotto ulteriori sviluppi, come i sistemi di *automated writing evaluation* (AWE) e le *computer-based assessments* (CBA), basati su algoritmi di NLP e tecniche di apprendimento automatico (Nardi, 2018; Tonelli et al., 2018). Tra questi strumenti, i chatbot conversazionali come ChatGPT, che si avvalgono di dispositivi di IAGen, stanno assumendo crescente rilevanza, sollevando un ampio dibattito scientifico sul loro potenziale e sui rischi connessi alla loro integrazione nei processi valutativi.

La letteratura mostra atteggiamenti ambivalenti degli insegnanti verso l'uso di ChatGPT nella valutazione. Alcuni studi ne evidenziano i vantaggi in termini di riduzione dei tempi di correzione e maggiore coerenza dei giudizi, soprattutto in contesti ad alto carico valutativo (Ruff et al., 2024; Zuckerman et al., 2023), nonché la possibilità di fornire feedback immediato agli studenti, permettendo ai docenti di concentrarsi su aspetti più complessi della scrittura (Pfau et al., 2023). Altri contributi segnalano invece criticità legate alla qualità del feedback e alla difficoltà dell'IA nel cogliere dimensioni soggettive della scrittura, quali originalità e costruzione argomentativa (Obata et al., 2024; Kim et al., 2024; Ljubojević et al., 2023), oltre al timore di una riduzione dell'autonomia professionale docente (Safrai et al., 2024). La letteratura sottolinea infine la necessità di una formazione specifica per un uso consapevole della GenAI, al fine di sviluppare competenze di mediazione tecnologica e ridurre disuguaglianze nelle pratiche valutative (Picciano et al., 2024; Jiang et al., 2024).

Muovendo da questa cornice teorica, il presente studio adotta un approccio esplorativo e interpretativo per indagare come gli insegnanti percepiscono l'inserimento di ChatGPT nella propria pratica valutativa. Il lavoro si colloca entro una concezione multifunzionale e formativa della valutazione, sviluppata in Italia da Giovannini (1994; 1995) e a livello internazionale da Broadfoot et al. (1999) attraverso il modello dell'*assessment for learning* (AfL). In questa prospettiva, la valutazione non è intesa come semplice misurazione degli esiti, ma come componente integrante del processo di apprendimento, orientata a generare feedback significativi per l'autoregolazione degli studenti e per la regolazione dell'agire didattico degli insegnanti (Shepard, 2000).

2. Metodo

Alla luce di tali riferimenti, l'indagine si propone di far emergere i punti di vista circa le opportunità, criticità, tensioni nonché i processi di rielaborazione soggettiva che si attivano quando le consolidate pratiche valutative degli insegnanti entrano in dialogo con dispositivi di intelligenza artificiale generativa (IAGen). L'obiettivo specifico è quello di esplorare i significati che gli insegnanti attribuiscono all'impiego di ChatGPT nella valutazione della produzione scritta degli studenti.

La ricerca adotta la forma dello *studio di caso* qualitativo a *orientamento esplorativo*, secondo l'impostazione proposta da Stake (1995). L'interesse quindi non è rivolto alla generalizzazione dei risultati ma alla comprensione contestuale delle esperienze individuali. I due casi presi in esame sono due docenti di italiano di scuola secondaria di primo grado selezionati in quanto portatori di posizioni iniziali differenti rispetto all'uso dell'IAGen: si tratta di un docente di circa quarant'anni, con cinque anni di esperienza e inizialmente più scettico rispetto all'impiego dell'IA (docente 1) e di una docente di circa cinquant'anni, con quindici anni di esperienza e una maggiore disponibilità a sperimentare nuove tecnologie (docente 2).

I dati sono stati raccolti attraverso *interviste in profondità* (Minichiello, Aroni & Hays, 2008) progettate per esplorare concezioni e percezioni in merito all'impiego di ChatGPT nella valutazione della produzione scritta degli studenti. L'intervista è stata preceduta da una fase preparatoria in cui ai docenti è stato richiesto, in primo luogo, di selezionare tre elaborati scritti rappresentativi di diversi livelli qualitativi (insufficiente, medio, ottimo) di una medesima classe, già corretti e valutati secondo i loro criteri abituali, e successivamente di compilare una scheda descrittiva per ciascuno studente. Gli elaborati selezionati, previo anonimizzazione rigorosa dei dati, sono stati successivamente analizzati con ChatGPT-4, dopo la trascrizione digitale dei testi e delle consegne. L'interazione con il chatbot è stata strutturata attraverso prompt specifici, elaborati sulla base della letteratura italiana sulla valutazione della scrittura con particolare riferimento alla griglia proposta da Calonghi (1972).

La scelta di adottare come riferimento la griglia di valutazione di Calonghi, anziché i criteri utilizzati individualmente dai docenti, è motivata dalla diversa formazione, esperienza professionale e modalità valutative degli insegnanti coinvolti. Al fine di garantire un quadro di riferimento condiviso e comparabile, si è pertanto optato per uno strumento standardizzato, conosciuto da entrambi i docenti.

Nello specifico, gli elementi presi in esame, così come articolati nella griglia

di Calonghi, hanno riguardato i contenuti, l'organizzazione del testo, la correttezza e la proprietà linguistica (punteggiatura, lessico, ortografia e grammatica). Il software, una volta forniti come riferimento la griglia di Calonghi (1972), il testo dell'elaborato, alcune informazioni socio-anagrafiche dello studente e la traccia proposta dal docente, ha prodotto le valutazioni articolandole secondo le macroaree previste dalla griglia stessa.

Terminata l'elaborazione da parte del software, i docenti sono stati intervistati separatamente mediante una traccia semistrutturata articolata in tre sezioni: (1) esplorazione delle conoscenze pregresse e degli atteggiamenti verso l'IA; (2) analisi degli output generati da ChatGPT e confronto con le loro valutazioni, avvenute invece in maniera tradizionale, come da consuetudine e abitudine dei docenti; (3) riflessioni sull'efficacia, sui limiti e sulle possibili implicazioni dell'uso del chatbot come supporto valutativo. Questa struttura ha permesso di cogliere sia i vissuti soggettivi sia le riflessioni critiche emerse dal confronto tra valutazione umana e automatizzata.

Le interviste sono state trascritte integralmente (verbatim) dopo aver ottenuto il consenso informato, garantendo il pieno rispetto della normativa vigente in materia di protezione dei dati personali. L'analisi ha seguito un approccio descrittivo-tematico per l'identificazione dei temi emergenti (ad esempio la conoscenza generale del software, l'approccio alla valutazione, i punti di forza e di criticità dell'impiego del software) attraverso una codifica iniziale mista (induttiva e deduttiva). Le categorie preliminari sono state integrate da codici generati durante la lettura approfondita delle trascrizioni. L'attenzione analitica è stata rivolta alla descrizione articolata dei punti di vista dei docenti, evitando eccessi di astrazione teorica o inferenze su strutture latenti non supportate dai dati.

3. Risultati

I risultati sono presentati seguendo l'articolazione dell'intervista. Per quanto riguarda la *familiarità con il software*, entrambi gli insegnanti hanno dichiarato di conoscere in modo generale strumenti come ChatGPT, pur senza averne fatto uso diretto. Il docente 1 ha espresso fin da subito una posizione più scettica, evidenziando il rischio di un impiego improprio da parte degli studenti: *«si usano a sproposito, per cose per cui non ci sarebbe bisogno… e si perdono occasioni per allenarsi e imparare»*. Il docente 2, pur riconoscendo la propria limitata esperienza, ha manifestato un atteggiamento più aperto, sostenendo che questi strumenti possano risultare utili se si conoscono le logiche d'uso.

Entrambi i docenti hanno inoltre descritto il proprio approccio valutativo

tradizionale, fondato su criteri strutturati quali correttezza ortografica, coerenza ed esattezza contenutistica e qualità della costruzione sintattica.

Il *confronto* tra le valutazioni dei docenti e quelle prodotte da ChatGPT ha evidenziato una serie di criticità negli output del chatbot. I docenti hanno rilevato una certa ripetitività nella rilevazione degli errori, la confusione tra diverse dimensioni linguistiche (morfologia, grammatica, sintassi) e la tendenza del software a classificare come problematici elementi perfettamente pertinenti al contesto, come nomi propri locali o espressioni idiomatiche. È stata inoltre segnalata la natura spesso generica dei feedback prodotti, percepiti come poco contestualizzati rispetto alle caratteristiche individuali degli studenti. Accanto a tali limiti, i docenti hanno riconosciuto alcuni *punti di forza*, in particolare, la capacità di ChatGPT di individuare errori "minimali", come ad esempio refusi, inversione di alcune lettere all'interno delle parole (pultroppo, sopratuto), che a volte sfuggono a una prima lettura umana e la rapidità con cui il chatbot valuta la pertinenza dell'elaborato rispetto alla traccia, contribuendo a ridurre il carico cognitivo richiesto al docente nelle fasi iniziali di analisi. Le criticità e le potenzialità individuate dai docenti possono essere ricondotte alle macroaree della griglia di Calonghi, che consentono di leggere in modo analitico il funzionamento del software. ChatGPT risulta maggiormente efficace nella dimensione della correttezza e proprietà linguistica, in cui l'analisi si fonda su elementi formalizzabili e ripetitivi, mentre incontra maggiori difficoltà nella valutazione dei contenuti, in particolare rispetto a originalità, stile ed espressività.

È emerso, infine, un disallineamento nei voti. Per gli elaborati di livello alto, i docenti hanno assegnato voti superiori rispetto all'IA; al contrario, per gli elaborati di livello basso, l'IA è stata più generosa dei docenti. Il "docente 1" ha interpretato il disallineamento come indice della mancanza, da parte dell'IA di "intelligenza pedagogica", evidenziando come essa conduca a una "correzione senza comprensione" incapace di cogliere stile, espressività e originalità e quindi poco sensibile agli elaborati di alto livello. Il "docente 2", invece, ha letto lo stesso fenomeno alla luce della funzione comunicativa del voto, sottolineando che *«Io il voto lo uso per comunicare qualcosa al ragazzo, non è solo una somma di errori»*, e spiegando così perché l'IA risulti più generosa con i testi deboli, standardizzando la valutazione verso il centro e non utilizzando il giudizio come segnale formativo intenzionale né in senso correttivo né valorizzante.

Nella parte conclusiva dell'intervista, entrambi gli insegnanti hanno espresso una *valutazione complessivamente* prudente ma interlocutoria dello strumento. ChatGPT è stato considerato potenzialmente utile come supporto, purché non sostituisca il giudizio professionale. Il docente 1 ha ipotizzato una possibile utilità *«nella formazione dei docenti inesperti»*, come occasione di confronto con un mo-

dello esterno che favorisca la riflessione critica sui propri criteri valutativi. Sono tuttavia emersi limiti significativi, tra cui la necessità, per gli insegnanti, di una formazione specifica per un uso consapevole dell'IAGen e l'incapacità del software di integrare variabili contestuali importanti, come il background culturale e linguistico degli studenti, elementi imprescindibili nel processo valutativo educativo.

4. Discussione e conclusione

In coerenza con l'approccio qualitativo e interpretativo adottato, l'attenzione non si è concentrata sull'efficacia tecnica dello strumento, ma sui processi di attribuzione di significato attivati dai partecipanti. Il confronto tra giudizi umani e valutazioni automatizzate ha evidenziato la coesistenza di atteggiamenti contrastanti – apertura e resistenza, curiosità e cautela, fiducia e diffidenza – che riflettono la complessità dell'incontro tra pratiche valutative consolidate e sistemi di nuova generazione. Una delle evidenze centrali emerse riguarda la difficoltà di riconoscere valore pedagogico a un dispositivo che, pur mostrando una certa coerenza nell'analisi formale, non è in grado di cogliere la dimensione semantica, stilistica e relazionale che caratterizza la scrittura scolastica. Si tratta di una resistenza che rimanda a quelle 'barriere di secondo ordine' (Ertmer, 1999), radicate nelle convinzioni profonde del docente sul senso dell'educare. La valutazione emerge così come un'operazione intrinsecamente interpretativa e contestuale, confermando la prospettiva situata descritta da Huot (2009), e appare difficilmente riducibile a un set di regole computazionali. Nel dialogo con i docenti, ChatGPT è apparso potenzialmente utile come supporto, ma solo se collocato all'interno di una cornice riflessiva che ne delimiti la funzione e ne problematizzi l'uso. Per gli insegnanti intervistati, dunque, l'obiettivo non è escludere a priori tali strumenti dalla pratica valutativa, bensì interrogare criticamente le condizioni e le finalità del loro impiego, alla luce di una concezione formativa e dialogica della valutazione (Giovannini 1994; 1995).

Dal punto di vista *metodologico*, lo studio conferma la pertinenza di un approccio qualitativo orientato all'ascolto delle voci degli insegnanti, capace di far emergere non solo dichiarazioni di convinzioni ma anche relative a resistenze, strategie di adattamento e tentativi di rielaborazione soggettiva, in cui i docenti rinegoziano il significato e il ruolo della tecnologia all'interno del proprio quadro professionale. Tale prospettiva si rivela particolarmente feconda in un ambito, quello dell'automazione della valutazione, dove dimensioni tecniche, etiche e pedagogiche si intrecciano in modo complesso e non eludibile.

I risultati indicano che gli insegnanti riconoscono a ChatGPT un'efficacia nell'analisi rapida e coerente degli aspetti formali del testo scritto, ma ne evidenziano i limiti nel cogliere le dimensioni semantica, pragmatica e contestuale proprie del giudizio professionale, inteso come processo interpretativo e relazionale. In questa prospettiva, l'IA può svolgere una funzione di supporto alla valutazione, senza sostituire il ruolo del docente come mediatore del significato, rendendo necessario un approccio ibrido che integri automazione e sensibilità pedagogica.

Lo studio presenta limiti legati al numero ridotto di partecipanti, alla necessità di trascrivere testi manoscritti per l'analisi automatizzata e al carico cognitivo richiesto dalla struttura dell'intervista. Tali elementi suggeriscono la necessità di perfezionare le strategie metodologiche e operative e di approfondire, in future ricerche, modalità di integrazione dell'IA più sostenibili e percorsi di formazione che ne favoriscano un uso critico e consapevole in ambito scolastico.

Riferimenti bibliografici

Broadfoot, P. M., Daugherty, R., Gardner, J., Gipp, C. V., Harlen, W., James, M., & Stobart, G. (1999). *Assessment for learning: Beyond the black box*. University of Cambridge School of Education.

Calonghi, L. (1972). *Valutazione delle composizioni scritte. Indicazioni docimologiche e psicometriche pratiche*. Roma: Armando.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), p. 319–340.

Giovannini, M. L. (1994). *Valutazione sotto esame*. Milano: Ethel editoriale Giorgio Mondadori.

Giovannini, M. L. (1995). *La valutazione: ovvero, oltre il giudizio sull'alunno*. Milano: Ethel Editoriale Giorgio Mondadori.

Goralski, M., & Tan, T. (2020). Artificial intelligence and sustainable development. *The International journal of Managment Education*.

Ertmer, P. A. (1999). Addressing first-and second-order barriers to change: Strategies for technology integration. *Educational Technology Research and Development*, 47(4), p. 47–61.

Huot, B. (2002). *Rearticulating writing assessment for teaching and learning*. Logan: University Press of Colorado.

Huot, B. A., & O'Neill, P. (Eds.) (2009). *Assessing writing: A critical sourcebook*. Boston: Bedford/St. Martin's.

Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? *Computers and Education*, p. 1-14.

Kim, H., Yang, J., Chang, D., & Lenke, L. (2024). Assessing the Reproducibility of the Structured Abstracts Generated by ChatGPT and Bard Compared to Human-Written Abstracts in the Field of Spine Surgery: Comparative Analysis. *Journal of Medical Internet Research*, 26, e52001.

Ljubojevi , D., Kadijevich, D., & Gutvajn, N. (2023). Towards a Valid and Reliable Checklist to Evaluate Argumentative Essays Composed by ChatGPT. *CEUR Workshop Proceedings* , 3696, p. 82-90.

Luckin, R. (2018). *Machine Learning and Human Intelligence: The Future of Education for the 21st Century*. London, UCL IOE Press.

Minichiello, V., Aroni, R., & Hays, T. (2008). *In-depth Interviewing: Principles, Techniques, Analysis*. (Third ed.) Camberwell: Pearson Australia Group.

Nardi, A. (2018). Valutare l'apprendimento online: una rassegna degli studi sull'e-testing nel contesto universitario. *Form@ re-Open Journal per la formazione in rete*, 18(1), p. 179-191.

Obata, A., Tagawa, T., & Ono, Y. (2023). Assessment of ChatGPT's Validity in Scoring Essays by Foreign Language Learners of Japanese and English. *Proceedings - 2023 15th International Congress on Advanced Applied Informatics Winter, IIAI-AAI-Winter 2023*, p. 105-110.

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, 2(3).

Picciano, A. (2024). Graduate Teacher Education Students Use and Evaluate ChatGPT as an Essay-Writing Tool. *Online Learning Journal*, 28(2).

Ruff, E., Engen, M., & Franz, J. (2024). ChatGPT Writing Assistance and Evaluation Assignments Across the Chemistry Curriculum. *Journal of Chemical Education*, 101(6), p. 2483-2492.

Safrai, M., & Orwig, K. (2024). Utilizing artificial intelligence in academic writing: an in-depth evaluation of a scientific review on fertility preservation written by ChatGPT-4. *Journal of Assisted Reproduction and Genetics*, 41(7), p. 1871–1880.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), p. 4-14.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks: SAGE Publications.

Tonelli, D., Grion, V., & Serbati, A. (2018). L'efficace interazione tra valutazione e tecnologie: evidenze da una rassegna sistematica della letteratura. *Italian Journal of Educational Technology*, p. 6-23.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), p. 425–478.

Wiliam, D. (2011). *Embedded Formative Assessment*. Bloomington: Solution Tree Press.

Zuckerman, M., Flood, R., & Tan, R. (2023). ChatGPT for assessment writing. *Medical Teacher*, 45 (11), p. 1224-1227

II.12
Rubriche IA-assistite per la valutazione formativa nella scuola secondaria di II grado: argomentazione di validità, feed-forward e audit di equità in chiave DD/UDL

Biagio Di Liberto
Cultore della materia di PAED-02/A, CeDisMa - Centro studi e ricerche sulla disabilità e marginalità
Dipartimento di Pedagogia, Università Cattolica del Sacro Cuore
biagio.diliberto@unicatt.it

Abstract

Lo studio presenta un modello integrato per l'uso copilotico dell'Intelligenza Artificiale (IA) nella progettazione e nell'impiego di rubriche valutative orientate all'Assessment for Learning in una classe terza di istituto tecnico. Il modello combina: rubriche IA-assistite con declinazioni differenziate in prospettiva DD/UDL; un Protocollo di Moderazione Valutativa Avanzato (PMVA) per garantire accordo inter-correttore; un audit di equità tramite Scheda DIF estesa a livello di descrittore. In un disegno a metodi misti, in linea con la Design-Based Research (Anderson, Shattuck, 2012), sono stati raccolti indicatori di efficienza (tempi di prototipazione e latenza del feedback), di qualità docimologica (coerenza criteri-livelli, κw) e di equità tra gruppi BES (L2, DSA, PEI). I risultati evidenziano una riduzione dei tempi di lavoro, un incremento dell'affidabilità inter-correttore e una riduzione dei differenziali valutativi per i sottogruppi più esposti a barriere. Il contributo discute le condizioni di validità d'uso delle rubriche IA-assistite entro un quadro di valutazione formativa inclusiva.

Parole chiave: rubriche IA-assistite; valutazione formativa; argomentazione di validità; equità valutativa; Differenziazione Didattica; Universal Design for Learning.

This paper presents an integrated model for the co-piloted use of AI in the design and implementation of rubric-based Assessment for Learning in a Grade 11 class in a technical upper secondary school. The model combines AI-assisted rubrics, differentiated in line with Differentiated Instruction and Universal Design for Learning (DI/UDL), an Advanced Assessment Moderation Protocol (PMVA) to support inter-rater agreement, and a descriptor-level equity audit based on extended DIF procedures. Within a mixed-methods, design-based research framework (Anderson, Shattuck, 2012), we collected indicators of efficiency

(rubric prototyping time, feedback latency), measurement quality (criteria–level coherence, weighted) and equity among SEN subgroups (L2 learners, students with specific learning disorders, students with individualized education plans). Findings indicate reduced processing time, increased inter-rater reliability and narrower assessment gaps for student groups most exposed to learning barriers. The paper discusses conditions for the valid and equitable use of AI-assisted rubrics within an inclusive formative assessment framework.

Keywords: AI-assisted rubrics; formative assessment; assessment equity; Differentiated Instruction; Universal Design for Learning; validity argument.

1. Introduzione

L'emergere dei Large Language Models (LLM) e degli strumenti generativi ha reso possibile la prototipazione rapida di rubriche valutative e la produzione quasi istantanea di feedback strutturati (Holmes, Bialik, Fadel, 2019; Miao, Holmes, 2023; Bender, Gebru, McMillan-Major, Shmitchell, 2021). La letteratura sui sistemi rubric-based e sullo scoring automatizzato richiama tuttavia cautele stringenti in termini di validità, trasparenza, equità tra sottogruppi e controllo professionale delle decisioni valutative (Williamson, Xi, Breyer, 2012; Zhang, D.-W., Boey, Tan, Jia, 2024; Zhang, Y., Nagarajan, Correnti, Boyatzis, 2024).

In chiave pedagogica, l'Assessment for Learning (AfL) valorizza rubriche che esplicitano criteri e standard, sostengono la feedback literacy e orientano il feed-forward, promuovendo l'autoregolazione anziché la mera certificazione dell'esito (Black, Wiliam, 1998; Zimmerman, 2002; Grion, Restiglian, 2019; Panadero, Jonsson, 2013). In questa cornice, la nozione di feedback spirals (Carless, 2019) supera la visione lineare del feedback come restituzione puntuale: le spirali di feedback implicano processi iterativi e ricorsivi in cui gli studenti elaborano informazioni da fonti multiple nel tempo, rivedendo non solo la prestazione ma anche le proprie strategie di apprendimento. Tale prospettiva si integra con lo sviluppo della feedback literacy (Carless, Boud, 2018) e con la valutazione sostenibile (Boud, Soler, 2016).

In questa prospettiva, la rubrica costituisce un artefatto che sostiene lo sviluppo del giudizio valutativo ai fini regolativi. Assumendo la prospettiva argomentativa della validità, l'introduzione dell'IA nei processi valutativi richiede evidenze non solo di coerenza interna e affidabilità, ma anche di conseguenze ed equità d'uso, con attenzione agli studenti L2 e con bisogni educativi speciali

(Messick, 1989; Kane, 2013; Mislevy, 2018; Williamson, Xi, Breyer, 2012; Miao, Holmes, 2023).

In ambito europeo, il recente quadro DigComp 3.0 (Cosgrove, Cachia, 2024) integra sistematicamente competenze di IA attraverso tutte le aree, enfatizzando la capacità di riconoscere bias algoritmici, valutare criticamente gli output e garantire equità nei processi valutativi digitali. L'integrazione fra Differenziazione Didattica (DD) e Universal Design for Learning (UDL) offre un quadro per progettare opportunità equipollenti di accesso ed espressione del costrutto valutato (CAST, 2024; d'Alonzo, 2016; Tomlinson, 2022; Di Liberto, d'Alonzo, 2024).

Parallelamente, gli studi sul Differential Item Functioning (DIF) consentono di verificare se descrittori formalmente identici operino in modo ingiustamente diverso per gruppi di studenti a parità di abilità (Zumbo, 1999; Penfield, Camilli, 2007; Agresti, 2010; Crane, Gibbons, Jolley, van Belle, 2006). La ricerca qui presentata considera l'IA come "copilota" e non come sostituto del giudizio docente, indagando in che misura rubriche IA-assistite, integrate con PMVA e audit di equità, possano migliorare qualità, efficienza e equità valutativa in un contesto reale di scuola secondaria di II grado (Boud, Soler, 2016; Grion, Restiglian, 2019).

L'uso copilotico implica che l'IA generi bozze e proposte, mentre il team docente esercita il controllo critico attraverso la formulazione del prompt, la valutazione dell'output e la revisione iterativa. In questa prospettiva, la competenza di prompting – intesa come capacità di strutturare istruzioni precise per ottenere risposte pedagogicamente appropriate – diventa parte integrante della professionalità docente nella progettazione valutativa (Cain, 2024; Cosgrove, Cachia, 2024; Miao, Holmes, 2023).

Le domande di ricerca sono:

- RQ1) In che misura le rubriche IA-assistite, integrate con pratiche di DD/UDL, riducono tempi di progettazione e latenza del feedback mantenendo o migliorando la qualità docimologica?
- RQ2) In che misura esse favoriscono chiarezza dei criteri, feed-forward e percezione di agency da parte degli studenti?
- RQ3) Gli audit di equità a livello di descrittore (DIF) consentono di ridurre i differenziali di esito per gruppi BES/L2/DSA senza alterare il costrutto valutato (Zumbo, 1999; Penfield, Camilli, 2007)?

2. Metodo

Lo studio è stato condotto nell'a.s. 2024-25 in una classe terza di un istituto tecnico ad indirizzo "Informatica e Telecomunicazioni" (N=24; 21 maschi, 3 studentesse), situato nel Nord-Ovest d'Italia. L'istituto presenta un'alta incidenza di studenti con background migratorio e si caratterizza per una consolidata cultura progettuale orientata all'inclusione. La distribuzione dei partecipanti comprendeva: 13 studenti senza BES, 5 con PDP (DSA), 2 con PEI, 4 NAI con italiano L2. L'avvio delle discipline di indirizzo, a partire dal secondo biennio, ha potenziato la proposta di compiti autentici in classe. Dal secondo anno la classe ha lavorato con rubriche formative, pur non IA-assistite: tale familiarità ha rappresentato una condizione favorevole per l'introduzione del modello oggetto di studio. La scelta è coerente con un impianto di Design-Based Research (Anderson, Shattuck, 2012).

Il compito autentico di riferimento ("IoT per la sicurezza e il benessere del laboratorio") ha impegnato la classe per 10-12 ore nel primo quadrimestre, con articolazione in 4 gruppi eterogenei da 6 studenti. La rotazione dei ruoli è stata supervisionata dal team docente con funzione protettiva per gli studenti con BES e potenziante per tutti i cluster. Il prodotto finale consisteva in un prototipo IoT end-to-end corredato da dossier tecnico e pitch di presentazione.

Lo sviluppo dell'esperienza si è articolato in cinque fasi distribuite su otto settimane del primo quadrimestre, con intervento copilotico dell'IA a tre livelli incrementali. Nella prima fase (settimane 1-2), l'IA ha supportato la co-progettazione della rubrica attraverso il ciclo iterativo prompt-output-revisione descritto in 2.1. Nella seconda fase (settimana 3), la rubrica è stata presentata agli studenti e si è avviato il compito autentico con formazione dei gruppi eterogenei. La terza fase (settimana 5) ha previsto il checkpoint intermedio con prima applicazione della rubrica; in questa fase l'IA è intervenuta come secondo livello copilotico, generando bozze di feedback individualizzato che il team docente ha poi validato, integrato e personalizzato. La quarta fase (settimana 6) ha visto l'analisi DIF sui risultati del checkpoint: il terzo livello copilotico ha supportato la revisione dei descrittori problematici, sottoponendo all'IA le criticità emerse e generando proposte di riformulazione successivamente vagliate dal team. La quinta fase (settimana 8) ha concluso il percorso con la consegna finale, la seconda misurazione e la restituzione del feed-forward. La riduzione della latenza di feedback da 7,5 a 2,2 giorni è dunque esito della sinergia incrementale dei tre livelli copilotici, non della sola chiarezza della rubrica. Il compito è stato progettato in coerenza con le competenze di indirizzo previste dalle Linee guida per gli Istituti Tecnici (D.P.R. 88/2010), con particolare riferimento alle aree "Si-

stemi e reti" e "Tecnologie e progettazione di sistemi informatici e di telecomunicazioni".

La Figura 1 sintetizza i quattro cluster della classe coinvolti nella progettazione DD/UDL (senza BES/L2; PDP; PEI; NAI-L2) e le principali leve di personalizzazione e di differenziazione (assi contenuto/processo/prodotto; ingaggio/rappresentazione/azione-espressione).

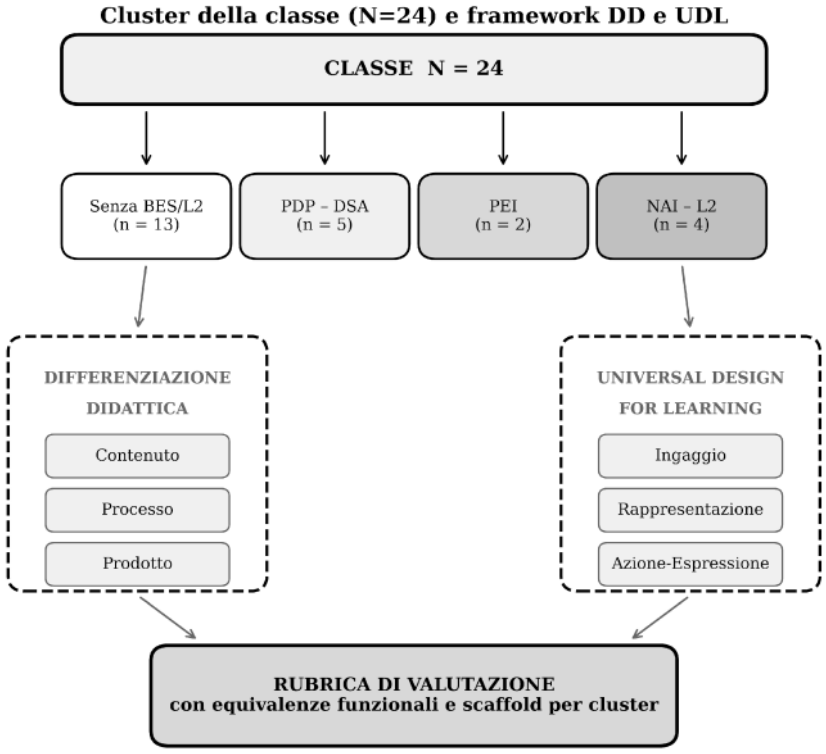


Fig. 1– Cluster della classe (N=24) e leve di differenziazione DD/UDL

Il team docente ha incluso: il docente di Telecomunicazioni (coordinatore), i docenti di Sistemi e Reti e TPSIT, l'ITP e il docente di sostegno, coinvolto sin dalla co-progettazione per la declinazione DD/UDL dei descrittori. Il docente di Italiano è stato coinvolto in una fase successiva, a seguito dell'analisi DIF, per la calibratura lessicale dei descrittori risultati problematici per il cluster L2. Il protocollo di ricerca è stato approvato dal Consiglio di classe e dal Dipartimento disciplinare.

Ciascun compito è stato declinato secondo principi DD/UDL, con requisiti

minimi espliciti e piste di estensione e altresì accomodamenti equivalenti (testi semplificati, scaffold grafici, opzioni multimodali di espressione, routine strutturate) (Castoldi, 2018, 2019; d'Alonzo et al., 2021; CAST, 2024; Tomlinson, 2022).

2.1 Disegno di ricerca e strumenti

Lo studio adotta un disegno a metodi misti, in linea con la Design-Based Research. La Figura 2 sintetizza il workflow in cinque fasi: prototipazione IA della rubrica, moderazione (PMVA), audit di equità (Scheda DIF estesa), analisi statistiche e confronto tra gruppi, feed-forward e ottimizzazione.

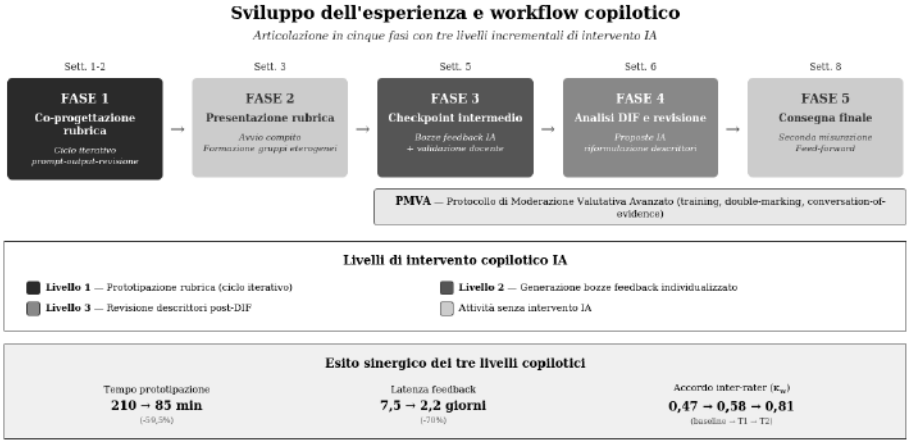


Fig. 2 – Sviluppo dell'esperienza e workflow copilotico. Articolazione in cinque fasi (8 settimane) con tre livelli incrementali di intervento IA e applicazione del PMVA a partire dal checkpoint intermedio. Dati: tempo di prototipazione 210→85 min (-59,5%); latenza feedback 7,5→2,2 giorni (-70%); κw 0,47→0,58→0,81 (baseline→T1→T2)

La rubrica IA-assistita prevede 8 criteri × 4 livelli (0-3): 1) giustificazione delle scelte progettuali; 2) uso di evidenze e dati; 3) progettazione tecnica e affidabilità; 4) comunicazione multimodale; 5) iniziativa e agency progettuale; 6) collaborazione e gestione del progetto; 7) etica, sicurezza e sostenibilità; 8) metacognizione e feed-forward.

Per ogni criterio sono esplicitati descrittori osservabili, esempi di evidenza e feed-forward (Jonsson, Svingby, 2007; Panadero, Jonsson, 2013; Brookhart, 2018). Tre campi trasversali incorporano DD/UDL nella rubrica: Modalità

equivalenti, Scaffold e DD e Barriera mirata, che specificano rispettivamente: forme equivalenti di manifestazione del costrutto, aiuti didattici consentiti e barriere (linguistiche, esecutive, visuospaziali, attentive) da presidiare.

La rubrica è stata co-progettata attraverso un ciclo prompt-output-revisione articolato in tre iterazioni che ha coinvolto l'LLM Gemini e il team docente:

Nella prima iterazione, il prompt iniziale ha specificato: la descrizione del compito autentico, le competenze di indirizzo target, la struttura richiesta (8 criteri × 4 livelli, scala 0-3) e il vincolo di declinazione DD/UDL per ciascun criterio. L'output grezzo presentava descrittori generici e scarsa differenziazione tra livelli adiacenti.

Nella seconda iterazione, il prompt è stato arricchito con esempi di evidenze attese per ciascun livello e con l'indicazione esplicita delle barriere da presidiare per i cluster L2, DSA e PEI. L'output ha mostrato maggiore specificità nei descrittori, ma persistevano sovrapposizioni tra i livelli 1 e 2.

Nella terza iterazione, il prompt ha incluso boundary samples (esempi di confine tra livelli) e la richiesta esplicita di formulare descrittori in forma osservabile e verificabile. L'output finale è stato sottoposto a revisione collegiale: il team tecnico (Telecomunicazioni, Sistemi e Reti, TPSIT) ha verificato la coerenza tra criteri e competenze e la progressione dei livelli; il docente di sostegno ha curato l'adeguatezza dei campi DD/UDL.

Il tempo complessivo di prototipazione (85 minuti, contro i 210 stimati per l'elaborazione manuale, -59,5%) include la formulazione dei prompt, l'analisi critica degli output e le micro-revisioni iterative, evidenziando che l'efficienza dipende dalla qualità del prompting e non dalla mera delega all'IA (Cain, 2024).

Criterio	Livello 0 - Iniziale	Livello 1 - Base	Livello 2 - Intermedio	Livello 3 - Avanzato
1. Giustificazione delle scelte progettuali	Scelte non motivate o incoerenti	Motivazioni parziali, legate a vincoli esterni	Giustificazioni coerenti con requisiti e risorse	Argomentazione articolata con analisi trade-off
2. Uso di evidenze e dati	Assenza di dati o uso scorretto	Dati presenti ma non interpretati	Interpretazione coerente dei dati	Integrazione critica di fonti multiple
3. Progettazione tecnica e affidabilità	Prototipo non funzionante o incompleto	Funzionamento parziale con errori	Funzionamento affidabile in condizioni standard	Gestione avanzata di eccezioni e stress-test
4. Comunicazione multimodale	Comunicazione assente o inadeguata	Presentazione comprensibile ma disorganizzata	Comunicazione chiara e adeguata al destinatario	Comunicazione efficace con integrazione multimodale

5. Iniziativa e agency progettuale	Esecuzione passiva di consegne	Prime scelte autonome su contenuto/ processo	Scelte motivate in relazione a readiness e interesse	Agency progettuale con assunzione di responsabilità
6. Collaborazione e gestione del progetto	Lavoro individuale senza coordinamento	Collaborazione episodica su sollecitazione	Coordinamento efficace con divisione ruoli	Leadership distribuita e gestione proattiva
7. Etica, sicurezza e sostenibilità	Aspetti etici non considerati	Consapevolezza parziale delle implicazioni	Applicazione di principi etici e di sicurezza	Analisi critica delle implicazioni sistemiche
8. Metacognizione e feed-forward	Assenza di riflessione sul processo	Riconoscimento di punti di forza/debolezza	Azioni di feed-forward esplicitate	Piano di miglioramento autonomo e realistico

Tab. 1 – Rubrica IA-assistita per il compito autentico "IoT per la sicurezza e il benessere del laboratorio" (8 criteri × 4 livelli)

Cluster	Modalità equivalenti	Scaffold e DD	Barriera mirata
L2 (NAI)	Pitch con supporto scritto visibile; interpretazione orale con grafico; tempi estesi	Glossario tecnico bilingue; sentence starters; mappe concettuali	Carico lessicale; fluenza orale; argomentazione scritta
DSA (PDP)	Registrazione audio per riflessione; scheda a scelta multipla con commento	Checklist debugging; step discreti; organizzatori grafici	Produzione scritta riflessiva; sequenzialità multi-step
PEI	Checklist etica con supporto visivo; discussione guidata; ruolo operativo valorizzato	Colloquio guidato con domande chiuse; routine strutturate; affiancamento peer	Astrazione; generalizzazione; introspezione autonoma

Tab. 2 – Declinazione DD/UDL della rubrica per cluster (sintesi)

Il disegno ha previsto due misurazioni all'interno dello stesso compito autentico: un checkpoint intermedio (al termine della fase di implementazione del prototipo) e la consegna finale. L'analisi DIF è stata condotta dopo la prima misurazione, consentendo la revisione dei descrittori problematici prima della seconda rilevazione.

Il Protocollo di Moderazione Valutativa Avanzato (PMVA) è stato applicato dopo la prima misurazione e ha incluso: training dei valutatori su boundary samples; standard setting ibrido; double-marking (2 valutatori × 4 artefatti, pari all'intera produzione dei gruppi); conversation-of-evidence con il team docente completo per la gestione argomentata dei disaccordi; monitoraggio del drift di severità/clemenza; documentazione delle decisioni (Brookhart, 2018; Boud & Soler, 2016).

L'audit di equità è condotto tramite una Scheda DIF estesa a livello di descrittore: per ciascun descrittore sensibile si registrano il gruppo a rischio, ipotesi di bias, esiti della regressione logistica ordinale e decisioni di revisione. Il focus è il testo del descrittore, non il singolo studente: eventuali segnali di DIF conducono alla riscrittura o all'aggiunta di supporti DD/UDL, non all'alterazione retrospettiva dei giudizi (Zumbo, 1999; Penfield & Camilli, 2007; Agresti, 2010; Crane et al., 2006).

Criterio	Liv.	Cluster	Ipotesi di bias	Esito	Intervento
2. Uso evidenze	2	L2	Il descrittore richiede lessico statistico e argomentazione scritta, misurando indirettamente la competenza linguistica	DIF moderato	Interpretazione orale con supporto grafico
3. Progettazione	3	DSA	Il livello avanzato richiede gestione simultanea di variabili e sequenze complesse, sovraccaricando le funzioni esecutive	DIF lieve	Mantenuto con scaffold: checklist; step discreti
4. Comunicazione	2	L2	Il descrittore valuta fluenza orale e ricchezza lessicale, penalizzando la competenza linguistica in sviluppo	DIF uniforme	Pitch con supporto scritto visibile; tempi estesi
5. Iniziativa	1	PEI	Il descrittore presuppone introspezione metacognitiva autonoma, difficoltosa per fragilità nella consapevolezza di sé	DIF moderato	Colloquio guidato con domande chiuse
7. Etica	3	PEI	Il livello avanzato richiede astrazione e generalizzazione di principi a contesti non familiari, penalizzando il pensiero concreto	DIF uniforme	Checklist etica con supporto visivo; discussione guidata
8. Metacognizione	2	DSA	Il descrittore richiede produzione scritta riflessiva con organizzazione testuale autonoma, sovraccaricando la memoria di lavoro	DIF moderato	Registrazione audio; scheda a scelta multipla

Tab. 3 – Scheda DIF estesa: analisi a livello di descrittore per cluster (ordinata per criterio)
Legenda: *DIF uniforme* = differenziale sistematico, richiede riscrittura o modalità alternative; *DIF moderato* = differenziale apprezzabile, richiede intervento; *DIF lieve* = differenziale contenuto, gestibile con scaffold minimo.

Le principali variabili quantitative sono: efficienza (tempi di prototipazione e latenza del feedback), qualità docimologica (coerenza criteri-livelli, w), equità (differenziali medi di punteggio tra cluster, indicatori di DIF) e percezione studentesca (chiarezza dei criteri, utilità del feedback). I dati qualitativi derivano da note di moderazione, verbali PMVA e risposte aperte in ottica di triangolazione mixed methods.

Nel loro insieme, questi elementi costituiscono il modello integrato delle rubriche IA-assistite, del PMVA e dell'audit di equità a livello di descrittore (Scheda DIF estesa) rappresentato nella Fig. 3.

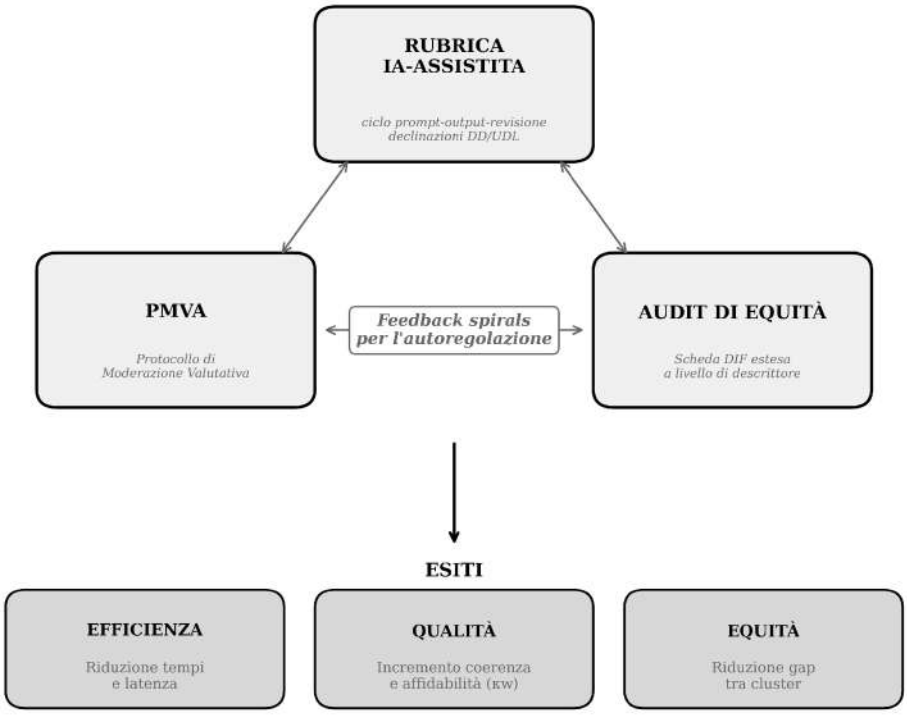


Fig. 3 – Modello integrato delle rubriche IA-assistite, del PMVA e dell'audit di equità a livello di descrittore

3. Risultati

Efficienza e qualità docimologica

Sul versante dell'efficienza, i dati attestano una riduzione del tempo di prototipazione della rubrica da 210 a 85 minuti (-59,5%) e una diminuzione della latenza media del feedback da 7,5 a 2,2 giorni (T-70%), a parità di numero di

studenti e compiti corretti. La coerenza criteri-livelli valutata tramite checklist esperta passa da 4,1 a 6,9 su scala 0-8, mentre l'affidabilità inter-correttore (κw), pari a 0,47 nelle esperienze pregresse senza modello IA-assistito, è aumentata a 0,58 al checkpoint intermedio e a 0,81 alla consegna finale, evidenziando un miglioramento progressivo attribuibile sia alla strutturazione della rubrica e al PMVA, sia alla revisione dei descrittori a seguito dell'analisi DIF (moderazione collegiale su trasparenza e affidabilità) (Jonsson, Svingby, 2007; Brookhart, 2018; Panadero, Jonsson, 2013).

La Fig. 4 mostra l'andamento di questi indicatori di efficienza (tempi di prototipazione, latenza del feedback) e di qualità docimologica (κw) nei diversi cicli di implementazione, evidenziando il miglioramento associato all'adozione del modello integrato.

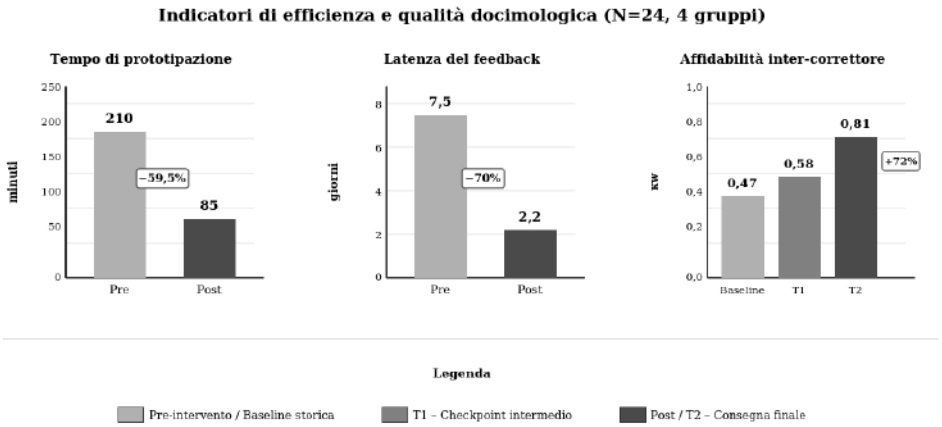


Fig. 4 – Indicatori di efficienza e qualità docimologica nei cicli di implementazione (N=24, 4 gruppi)

Equità tra cluster di studenti

In termini di equità, il divario medio di punteggio tra studenti L2/DSA e studenti senza BES, su scala rubrica 0-3, si riduce da 0,9 a 0,4 punti a seguito della revisione dei descrittori e dell'integrazione sistematica degli accomodamenti equivalenti DD/UDL. Tutti i cluster registrano un incremento dell'indice di rubrica e della chiarezza criteriale percepita, con progressi più marcati nei gruppi maggiormente esposti a barriere.

Le analisi ordinali a livello di descrittore individuano segnali di DIF uniforme a sfavore degli studenti L2 su descrittori ad alto carico lessicale o implicitamente ancorati a pratiche culturali specifiche. In tali casi, la fairness review qualitativa e la revisione mirata hanno determinato riscritture dei descrittori e un rafforza-

mento delle modalità equivalenti (esempi visivi, opzioni orali con supporti scritti), con attenuazione dei differenziali nelle iterazioni successive (Crane et al., 2006; Agresti, 2010).

La Figura 5 mostra la riduzione del divario tra studenti senza BES e i gruppi L2/DSA/PEI prima e dopo l'applicazione degli accomodamenti DD/UDL, evidenziando l'efficacia del modello in termini di equità tra cluster.

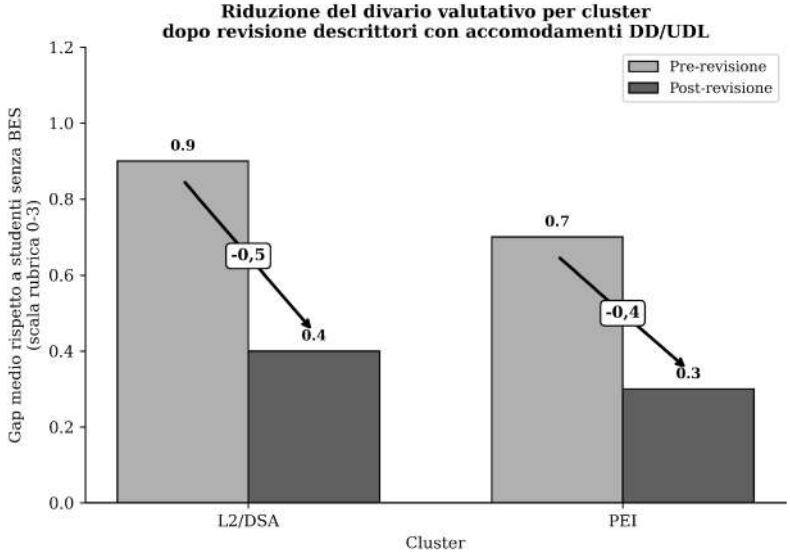


Fig. 5 – Riduzione del divario medio (scala 0-3) tra studenti senza BES e cluster L2/DSA/PEI, pre/post revisione DD/UDL

Percezioni degli studenti

I questionari rilevano un aumento della chiarezza dei criteri (media da 3,0 a 4,2 su 5) e un miglioramento della percezione della partecipazione. Alcuni studenti riferiscono che la rubrica "fa capire meglio cosa manca" e che il feed-forward per criterio consente di "sapere quale passo intraprendere", soprattutto nei compiti tecnici più complessi. Un ciclo sperimentale in cui le rubriche sono state utilizzate senza narrazioni esemplari e senza moderazione collegiale ha evidenziato un parziale appiattimento sui criteri di iniziativa e creatività, poi riassorbito riattivando le routine PMVA: ciò segnala il rischio di un uso riduttivo delle rubriche IA-assistite come dispositivi di standardizzazione.

4. Discussione

I risultati confermano che le rubriche IA-assistite, in un impianto di AfL con moderazione collegiale, producono miglioramenti simultanei di efficienza, qualità docimologica ed equità valutativa. Si rileva come l'incremento di κ_w e della coerenza criteri-livelli sia coerente con la letteratura sulle rubriche analitiche e sui benefici del training con esempi di confine e conversation-of-evidence (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Brookhart, 2018).

I risultati suggeriscono inoltre che l'efficacia delle rubriche IA-assistite dipenda in misura rilevante dalla competenza di prompting del team docente. La capacità di formulare prompt che esplicitino vincoli didattici (struttura, progressione tra livelli, inclusività), forniscano esempi di confine e richiedano output osservabili si configura come una competenza professionale emergente, coerente con il framework DigComp 3.0 (Cosgrove & Cachia, 2024) e con la letteratura sul prompt engineering in ambito educativo (Cain, 2024). L'assenza di tale competenza rischia di ridurre l'IA a generatore di contenuti generici, vanificando il potenziale copilotico e riproducendo le criticità dei sistemi di scoring automatizzato non presidiati.

L'integrazione esplicita di DD/UDL nei campi della rubrica ha consentito di trattare la personalizzazione non come intervento emergenziale legato ad etichettamenti (DSA, PEI, L2), ma come componente intrinseca e plurale del design valutativo, coerente con i principi di autonomia e competenza propri della Self-Determination Theory (Deci & Ryan, 2000). La riduzione del gap per L2/DSA suggerisce che accomodamenti "equivalenze funzionali" possano migliorare l'accesso ai livelli più alti senza compromettere il costrutto. Al tempo stesso, l'episodio di appiattimento sui criteri di iniziativa e creatività segnala che l'IA, se sganciata da cornici professionali robuste, può favorire una standardizzazione eccessiva, con effetti collaterali sulle dimensioni più aperte del compito. Ciò richiama in termini robusti le riflessioni sulla valutazione sostenibile (Boud & Soler, 2016; Grion & Restiglian, 2019).

5. Conclusioni, limiti e prospettive

Lo studio propone un modello operativo per l'uso copilotico dell'IA nella valutazione formativa di classe, radicato in cornici teoriche consolidate e dotato di un'argomentazione di validità che integra contenuto, processi di risposta, struttura interna, conseguenze ed equità di utilizzo (Messick, 1989; Kane, 2013; Mi-

slevy, 2018; Miao & Holmes, 2023). L'intreccio tra rubriche IA-assistite, PMVA e audit di equità a livello di descrittore dimostra la possibilità di coniugare efficienza, qualità docimologica e processi d'inclusione, a condizione che il controllo delle decisioni ad alto impatto resti in capo al team docente.

Lo studio è limitato a una singola classe; le analisi DIF sono condotte su sottogruppi ridotti e dal ruolo del docente-ricercatore. Lo studio si è concentrato su un singolo compito autentico ("IoT per la sicurezza e il benessere del laboratorio") che ha consentito due misurazioni all'interno dello stesso percorso. Gli altri compiti autentici plurimodali previsti nella progettazione di classe (infografica argomentata, recensione critica, demo tecnica) sono in corso di realizzazione e costituiscono prospettive di replicabilità del modello. La co-valutazione con gli studenti non è stata inclusa nel presente disegno per ragioni di complessità organizzativa e di protezione dei cluster più vulnerabili in fase di prima implementazione; tuttavia, forme leggere di coinvolgimento (retrospettiva di gruppo, checkpoint auto-valutativo) sono state presenti.

In prospettiva, sarà necessario verificare il modello in altri ordini di scuola e discipline, esplorando le condizioni di scalabilità e sostenibilità delle rubriche IA-assistite in cornice DD/UDL. Lo studio non ha indagato sistematicamente le strategie di prompting né le competenze pregresse dei docenti in questo ambito: future ricerche potranno esplorare la relazione tra qualità del prompt e qualità della rubrica generata, nonché i percorsi formativi necessari per sviluppare la competenza di prompting in chiave didattica (Cain, 2024). Future iterazioni potranno altresì integrare forme strutturate di peer assessment e auto-valutazione (Panadero & Jonsson, 2013; Zhang, Boey, Tan & Jia, 2024; Zhang et al., 2024; Bender et al., 2021).

Riferimenti bibliografici

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken, NJ: Wiley.

Anderson, T., Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher*, 41(1), 16-25. https://doi.org/-10.3102/0013189X11428813

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). ACM.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7-74.

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, 41(3), 400-413. https://doi.org/10.1080/02602938.-2015.1018133

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, 3, 22.

CAST. (2024). *Universal Design for Learning Guidelines version 3.0*. Retrieved July 30, 2024, from https://udlguidelines.cast.org

Cain, W. (2024). Prompting Change: Exploring Prompt Engineering in Large Language Model AI and Its Potential to Transform Education. *TechTrends*, 68, 47-57. https://doi.org/10.1007/s11528-023-00896-0

Carless, D. (2019). Feedback loops and the longer-term: Towards feedback spirals. *Assessment & Evaluation in Higher Education*, 44(5), 705-714. https://doi.org/-10.1080/02602938.2018.1531108

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Castoldi, M. (2018). *Compiti autentici. Un nuovo modo di insegnare e apprendere*. Torino: SEI.

Castoldi, M. (2019). *Valutare e certificare le competenze*. Roma: Carocci.

Cosgrove, J., & Cachia, R. (2024). *DigComp 3.0: European Digital Competence Framework – Fifth Edition*. Luxembourg: Publications Office of the European Union. https://data.europa.eu/doi/10.2760/0001149

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques (DIF-with-par). *Medical Care*, 44(11 Suppl 3), S115-S123.

d'Alonzo, L. (2016). *La differenziazione didattica per l'inclusione. Metodi, strategie, attività*. Trento: Erickson.

d'Alonzo, L., Bocci, F., & Pinnelli, S. (2021). *Didattica speciale per l'inclusione* (5a ed.). Brescia: Scholé.

Deci, E. L.,& Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227-268.

Di Liberto, B.,& d'Alonzo, L. (2024). Assessment for learning in "Differentiated Instruction". Authentic educational constructs and processes of inclusive protagonism. *QTimes – Journal of Education, Technology and Social Studies*, XVI(2). https://www.qtimes.it

Grion, V., & Restiglian, E. (2019). *Valutare per comprendere*. Milano: FrancoAngeli.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence in Education*. Boston, MA: Center for Curriculum Redesign.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. https://doi.org/10.1111/jedm.12000

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: American Council on Education/Macmillan.
Miao, F., & Holmes, W. (2023). *Guidance on generative AI in education and research*. Paris: UNESCO.
Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. New York: Routledge.
Panadero, E.,& Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129-144.
Penfield, R. D., & Camilli, G. (2007). Differential item functioning and fairness. In C. R. Rao, S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 125-167). Amsterdam: Elsevier.
Tomlinson, C. A. (2022). *La differenziazione didattica in classe. Per rispondere ai bisogni di tutti gli alunni* (a cura di L. d'Alonzo). Brescia: Morcelliana-Scholé.
Williamson, D. M., Xi, X.,& Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2-13. https://doi.org/10.1111/j.1745-3992.2011.00223.x
Zhang, D.-W., Boey, M., Tan, Y. Y., & Jia, A. H. S. (2024). Evaluating large language models for criterion-based grading from agreement to consistency. *npj Science of Learning*, 9, Article 79.
Zhang, Y., Nagarajan, A., Correnti, R., & Boyatzis, R. (2024). Generative AI and automated scoring of critical thinking: Validity evidence from human and machine ratings. *Journal of Educational Measurement*, 61(1), 56-78.
Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, 41(2), 64-70.
Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: DHRE.

II.13
Interazioni con ChatGPT e impatto sulle attività metacognitive degli studenti: uno studio sulla risoluzione di problemi matematici nella scuola secondaria di secondo grado

Umberto Dello Iacono
Università della Campania "L. Vanvitelli", umberto.delloiacono@unicampania.it
Marco Picariello
Università della Campania "L. Vanvitelli", marco.picariello@unicampania.it

Abstract

This exploratory study investigates the potential of ChatGPT to promote metacognitive activities in mathematical problem-solving contexts. Twelve female students in their final year of upper secondary school were involved in two laboratories: the first with non-configured ChatGPT, and the second in which the chatbot was configured through a structured prompt to behave as an educational tutor and encourage metacognitive activities of orientating, planning, executing, monitoring, evaluation, and elaboration. Interactions among students and with the chatbot were recorded, transcribed, and coded according to the taxonomy proposed by Meijer and colleagues. A qualitative analysis shows that, only with the structured prompt, ChatGPT is able to promote advanced metacognitive processes, highlighting the crucial role of dialogic design in the pedagogical effectiveness of the tool.

Keywords: mathematics education; chatbots; metacognition; problem solving; prompt engineering.

Il presente studio esplorativo indaga il potenziale di ChatGPT nel promuovere attività metacognitive in contesti di problem solving matematico. Dodici studentesse dell'ultimo anno della scuola secondaria di secondo grado sono state coinvolte in due laboratori: il primo con ChatGPT non configurato, e il secondo nel quale il chatbot è stato configurato attraverso un prompt strutturato per comportarsi come tutor educativo e favorire attività metacognitive di orientamento, pianificazione, esecuzione, monitoraggio, valutazione ed elaborazione. Le interazioni tra studenti e con il chatbot sono state registrate, trascritte e codificate secondo la tassonomia proposta da Meijer e colleghi. Un'analisi qualitativa mostra

che, solo con un prompt strutturato, ChatGPT è in grado di favorire processi metacognitivi avanzati, evidenziando il ruolo cruciale della progettazione dialogica nell'efficacia pedagogica dello strumento.

Parole chiave: didattica della matematica; chatbot; metacognizione; problem solving; prompt engineering.

1. Introduzione e quadro teorico

Molti studi recenti mostrano il potenziale dell'Intelligenza Artificiale (AI) generativa e dei Modelli Linguistici di Grandi dimensioni (LLMs), come ChatGPT, nell'insegnamento e apprendimento della matematica (ad es., Taani & Alabidi, 2025). In particolare, tali modelli offrono nuove opportunità per l'apprendimento personalizzato grazie a feedback immediati che favoriscono la riflessione e l'auto-spiegazione concettuale da parte degli studenti. Pepin et al. (2025) mettono in luce come ChatGPT possa costituire un valido supporto per gli studenti, sia nel consolidamento dei concetti matematici di base sia nell'affrontare contenuti più specifici della disciplina.

La possibilità di interagire con sistemi basati su LLMs, come ChatGPT, trasforma il rapporto tra studente e contenuto matematico: non è più mediato esclusivamente dal libro di testo, dall'insegnante o dalla discussione tra pari, ma anche da una fonte esterna capace di generare risposte articolate e contestualizzate. L'interazione con un chatbot non si limita all'accesso passivo all'informazione, ma assume la forma di una conversazione dinamica, riflessiva e regolativa dei processi cognitivi e metacognitivi. Tale scenario solleva interrogativi cruciali per la didattica della matematica relativamente allo sviluppo del pensiero matematico, al ruolo dell'insegnante e al ruolo del feedback.

Il presente studio si configura come un'indagine esplorativa di natura qualitativa e si focalizza sull'ultima questione, ossia sul ruolo del feedback. Ci siamo chiesti se ChatGPT, configurato per guidare il ragionamento e le riflessioni degli studenti, possa diventare un alleato metacognitivo, ossia favorire attività metacognitive.

2. Quadro concettuale

Un'attività metacognitiva è da intendersi "un'applicazione strategica della conoscenza metacognitiva per raggiungere obiettivi cognitivi" (Meijer et al., 2006,

p. 209, trad. nostra). Meijer et al. (2006) identificano sei categorie di attività metacognitive osservabili durante la risoluzione di compiti cognitivi complessi, che non seguono una sequenza rigida, ma che sono dinamicamente intrecciate:

- orientamento: attivare conoscenze pregresse, identificare requisiti e dati, formulare ipotesi, rileggere domande, osservare tabelle o diagrammi;
- pianificazione: identificare informazioni, definire sotto-obiettivi, riformulare, pensare a ritroso, formulare un piano, semplificare il problema;
- esecuzione: commentare, evidenziare, prendere appunti, rispondere a domande, stimare valori, attuare il piano;
- monitoraggio: dichiarare la comprensione (anche parziale), riconoscere la mancata comprensione, identificare e correggere gli errori, notare incongruenze;
- valutazione: controllare il lavoro svolto, spiegare e giustificare la strategia adottata, interpretare, esprimere incertezza sulle conclusioni, esercitare autocritica;
- elaborazione: trarre conclusioni, sintetizzare, commentare difficoltà, riflettere su abitudini personali.

La letteratura sulla metacognizione nel problem solving matematico evidenzia come gli studenti non esperti tendano a seguire percorsi lineari, talvolta procedurali e poco riflessivi, mentre gli studenti esperti alternano in maniera flessibile fasi di pianificazione, monitoraggio e valutazione (Schoenfeld, 2016). La tassonomia di Meijer et al. (2006) offre una cornice utile per analizzare tali dinamiche in modo empirico, poiché concepisce le fasi metacognitive come processi dinamici e riattivabili in base alle esigenze cognitive dello studente.

Contel e Cusi (2024) adattano la tassonomia proposta da Meijer et al. (2006) al contesto della didattica della matematica e, in particolare, alla risoluzione di problemi e mostrano come ChatGPT, senza una progettazione mirata dell'interazione, non è in grado di promuovere tutte le attività metacognitive secondo Meijer et al. (2006), in particolare la valutazione e l'elaborazione.

Alla luce di queste considerazioni, abbiamo progettato un'attività didattica che ha coinvolto studenti dell'ultimo anno della scuola secondaria di secondo grado. Abbiamo analizzato qualitativamente le interazioni tra studenti e ChatGPT durante attività di problem solving matematico con l'obiettivo di indagare in che misura ChatGPT, configurato tramite un prompt strutturato per agire come un tutor educativo, sia in grado di promuovere negli studenti attività di orientamento, pianificazione, esecuzione, monitoraggio, valutazione ed elaborazione.

3. Metodo

Lo studio è stato condotto con 12 studentesse di classe quinta di un Liceo delle Scienze Umane (di seguito partecipanti), che hanno lavorato in piccoli gruppi a due laboratori della durata di un'ora ciascuno. Nel primo laboratorio, ChatGPT è stato inizializzato tramite il seguente prompt non strutturato: "Sono una studentessa del quinto anno di scienze umane. Vorrei che tu fossi il mio tutor di matematica. Devo imparare le derivate".

Nel secondo laboratorio, ChatGPT è stato configurato tramite un prompt strutturato per comportarsi come tutor educativo, guidando il processo senza fornire direttamente le soluzioni, personalizzando le domande in base alle risposte degli studenti, stimolando il pensiero critico attraverso domande aperte, valorizzando gli sforzi anche in caso di errori e promuovendo le attività metacognitive. Il prompt ha incluso anche il compito iniziale per i partecipanti, ossia un problema reale relativo alla modellizzazione della crescita esponenziale degli iscritti a un liceo. Ogni gruppo ha avuto a disposizione un pc. I partecipanti potevano rileggere, scorrere, evidenziare parti del testo, discutere tra loro e interagire liberamente con il chatbot.

Abbiamo raccolto le registrazioni video degli schermi e l'audio dei dialoghi interni ai gruppi. Abbiamo trascritto le interazioni studente-studente e studente–ChatGPT nonché i comportamenti dei partecipanti (scroll, riletture, evidenziazioni). Ogni intervento è stato codificato secondo la tassonomia di Meijer et al. (2006), identificando per ciascun segmento l'attività metacognitiva prevalente. La codifica è stata condotta dai ricercatori a partire da una lettura condivisa dei dati. Le discrepanze sono state discusse fino al raggiungimento di un accordo.

4. Risultati

Durante il primo laboratorio, con prompt non strutturato, gli studenti si sono concentrati principalmente su attività di orientamento e pianificazione, con scarsa attenzione alle fasi di esecuzione, monitoraggio e valutazione, così come mostrato in Fig. 1 relativamente a uno dei gruppi coinvolti.

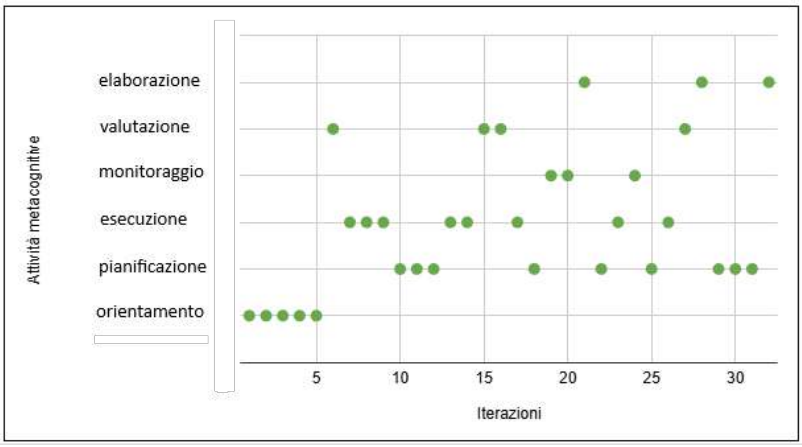


Fig. 1 – Attività metacognitive di uno dei gruppi che hanno interagito con ChatGPT con prompt non strutturato

L'elaborazione è comparsa solo nella parte finale del laboratorio. Le fasi di valutazione si sono manifestate episodicamente, e le conversazioni hanno mantenuto uno stile prevalentemente procedurale. Risultati analoghi hanno riguardato anche gli altri gruppi.

Al contrario, durante il secondo laboratorio, in cui ChatGPT è stato configurato tramite un prompt strutturato per guidare le attività metacognitive, il tempo trascorso tra le singole interazioni studente/studente o studente/ChatGPT durante la sessione con prompt strutturato è stato maggiore rispetto alla sessione di laboratorio precedente, così come mostrato in Fig. 2 relativamente a uno dei gruppi. Ciò sembra essere l'effetto di momenti di riflessione più lunghi tra una domanda e la risposta successiva da parte dei partecipanti a partire dagli stimoli offerti da ChatGPT durante la sessione con prompt strutturato rispetto alla precedente.

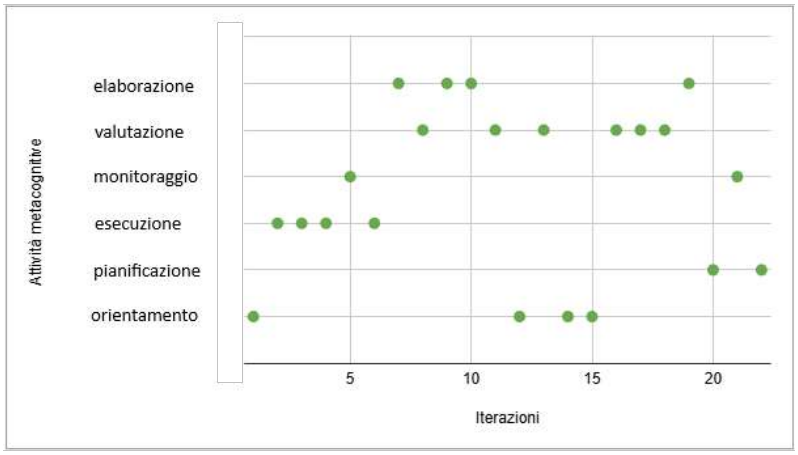


Fig. 2 – Attività metacognitive di uno dei gruppi che hanno interagito con ChatGPT con prompt strutturato

L'attività di elaborazione si sia innescata ben quattro volte e tre volte già nella prima metà dell'attività. La Fig. 2 mostra, inoltre, come il susseguirsi delle attività metacognitive innescate dal gruppo non sia stato lineare o sequenziale, come nel caso di prompt non strutturato: gli studenti sono saltati continuamente da un'attività metacognitiva all'altra e si sono coinvolti in tutte le attività senza lunghi momenti su una specifica attività.

In sintesi, l'analisi qualitativa dei dati raccolti sembra indicare una differenza marcata tra le due sessioni di laboratorio in termini attività metacognitive attivate. Nel primo laboratorio, con prompt non strutturato, l'interazione con ChatGPT è apparsa breve, lineare e finalizzata a ottenere un "passo successivo", senza approfondire la comprensione del modello né giustificare le scelte. Al contrario, durante il secondo laboratorio con prompt strutturato, i partecipanti hanno mostrato riflessioni più profonde e una maggiore regolazione dei processi cognitivi. L'interazione non è stata finalizzata a confermare i passaggi matematici, ma a interpretare il problema, discutere la plausibilità del modello, confrontare le rappresentazioni e giustificare le scelte. I partecipanti hanno riletto più volte le risposte del chatbot, hanno ripercorso il proprio ragionamento, ritornando sui dati iniziali e confrontando ipotesi alternative.

5. Discussione

Studi recenti sull'uso dei chatbot nella didattica della matematica hanno evidenziato due modalità principali di interazione: uno stile "procedurale", in cui il chatbot è consultato come fonte di conferma, e uno stile "riflessivo", in cui l'IA stimola discussione, giustificazione e revisione del ragionamento (Contel & Cusi, 2025; Filiz & Gür, 2025). La differenza non risiede nella tecnologia in sé, ma nell'"architettura dialogica" con cui tale tecnologia viene inserita nel contesto didattico.

I risultati di questo studio esplorativo suggeriscono come l'attivazione dei processi metacognitivi è fortemente influenzata dall'ingegnerizzazione del prompt (Knoth et al., 2024). I risultati del primo laboratorio confermano quanto osservato da Contel e Cusi (2025), mostrando una bassa frequenza di interventi di valutazione e quasi totale assenza di elaborazione da parte degli studenti durante un'attività nella quale hanno interagito con ChatGPT non configurato. Tuttavia, la seconda sessione di laboratorio ha evidenziato come ChatGPT, adeguatamente configurato, possa potenziare l'autonomia strategica degli studenti e favorire attività metacognitive negli studenti durante il problem solving matematico. Il numero complessivo di interazioni degli studenti è diminuito nel secondo laboratorio, suggerendo una riflessione più profonda e prolungata a partire dagli stimoli forniti dal chatbot. Il prompt strutturato ha favorito riflessioni metacognitive più profonde, orientando l'interazione all'interpretazione del problema, alla valutazione del modello e alla revisione critica del ragionamento. L'efficacia del laboratorio è mostrata anche dal susseguirsi delle attività metacognitive innescate dal gruppo in maniera non lineare (Schoenfeld, 2016). Gli studenti hanno saltato continuamente da un'attività metacognitiva all'altra e si sono coinvolti in tutte le attività metacognitive. La differenza osservata tra le due sessioni di laboratorio sembra dipendere dalla struttura dialogica introdotta dal prompt progettato. ChatGPT, adeguatamente configurato, si è rivelato efficace nel promuovere le fasi di valutazione ed elaborazione, trasformando l'interazione in uno spazio di riflessione metacognitiva più ricco e articolato. I dati sembrano indicare che ChatGPT non può essere considerato un tutor metacognitivo "per natura", ma può assumere tale ruolo se la sua interazione è progettata per promuovere la riflessione da parte degli studenti. Il prompt strutturato non agisce come vincolo rigido, ma come cornice regolativa che orienta lo studente a sviluppare una prospettiva più critica e consapevole sul problema matematico. In questo contesto, l'IA non fornisce soluzioni già pronte, ma apre un dialogo che sostiene l'analisi, la giustificazione e la valutazione delle scelte.

Il prompt assume il ruolo di strumento didattico, attraverso il quale l'inse-

gnante orchestra il dialogo e orienta lo studente verso pratiche riflessive. In questa prospettiva, ChatGPT non agisce come risolutore del problema, ma come mediatore metacognitivo a supporto dell'analisi, della giustificazione e della valutazione delle scelte.

In conclusione, lo studio mostra che ChatGPT è in grado sostenere attività metacognitive avanzate nella risoluzione di problemi matematici solo se l'interazione è strutturata con intenzionalità educativa. Ciò non implica che la tecnologia possieda intrinsecamente capacità didattiche, ma che la sua efficacia dipende dalla qualità pedagogica dell'orchestrazione dell'interazione. In ambito scolastico, ciò implica la necessità di utilizzare prompt strutturati come strumenti pedagogici per orchestrare il dialogo, promuovere processi metacognitivi e orientare gli studenti verso pratiche riflessive.

Risultano necessari studi futuri, con campioni più ampi e diversificati, per approfondire e validare le evidenze emerse.

Riferimenti bibliografici

Contel, F., & Cusi, A. (2025). Investigating the Role of ChatGPT in Supporting Metacognitive Processes During Problem-Solving Activities. *Digital Experiences in Mathematics Education*, 11, 167-191. https://doi.org/10.1007/s40751-024-00164-7

Filiz, A., & Gür, H. (2025). Students' Perceptions and Applications of Metacognitive Awareness Levels in Problem Solving with ChatGPT. *Educational Process: International Journal*, 14, e2025063. https://doi.org/10.22521/edupij.2025.14.63

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6, 100225.

Meijer, J., Veenman, M.,& van Hout-Wolters, B. (2006). Metacognitive activities in text-studying and problem-solving: Development of a taxonomy. *Educational Research and Evaluation*, 12(3), 209–237. https://doi.org/10.1080/138036105004-79991

Pepin, B., Buchholtz, N., & Salinas Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education*, 11(1), 9–41. https://doi.org/10.1007/s40751 025 00172-1

Schoenfeld, A. H. (2016). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics (Reprint). *Journal of education*, 196(2), 1-38.

Taani, O., & Alabidi, S. (2025). ChatGPT in education: Benefits and challenges of ChatGPT for mathematics and science teaching practices. *International Journal of Mathematical Education in Science and Technology*, 56(9), 1748-1777. https://doi.org/10.1080/0020739X.2024.2357341

II.14
Potenzialità e limiti nell'uso dell'approccio strumentale nell'interpretazione delle interazioni tra studenti e IA generativa

Francesco Contel
Sapienza Università di Roma, francesco.contel@uniroma1.it
Annalisa Cusi
Sapienza Università di Roma, annalisa.cusi@uniroma1.it

Abstract

Lo studio indaga l'uso della IA generativa come strumento di supporto metacognitivo in attività di problem solving matematico. Si esplorano le implicazioni teoriche dell'impiego del costrutto della genesi strumentale a partire da uno studio di caso esplorativo in cui si esaminano le interazioni di uno studente con GPT-4o. In particolare, si mostra come l'emergere di schemi d'uso orientati alla validazione sia condizionato dall'indeterminatezza del modello linguistico. L'analisi mostra come le componenti deterministiche degli schemi coesistano con forme di instabilità generate dalla GenAI, incidendo sulla coerenza del ragionamento matematico co-costruito. I risultati confermano l'utilità dell'approccio strumentale ma ne sottolineano i limiti nel descrivere le interazioni con strumenti di IA.

Parole chiave: IA generativa; ChatGPT; genesi strumentale; interazione umano-IA; didattica della matematica.

The study investigates the use of generative AI as a metacognitive support tool in mathematical problem solving activities. It explores the theoretical implications of using the instrumental genesis construct based on an exploratory case study examining a student's interactions with GPT-4o. In particular, it shows how the emergence of validation-oriented utilisation schemes is conditioned by the indeterminacy of the linguistic model. The analysis shows how the deterministic components of the schemes coexist with forms of instability generated by GenAI, affecting the consistency of the co-constructed mathematical reasoning. The results confirm the usefulness of the instrumental approach but highlight its limitations in describing interactions with AI tools.

Keywords: GenAI; ChatGPT; instrumental genesis; human-GenAI interaction; mathematics education.

Introduzione

Il carattere mutevole dei modelli linguistici alla base dell'intelligenza artificiale generativa (GenAI) pone alcuni problemi di ordine teorico e metodologico per i ricercatori in didattica della matematica: la GenAI non possiede le caratteristiche di una tecnologia educativa e i costrutti teorici elaborati in passato non sembrano sufficientemente flessibili per catturare la specificità delle interazioni umano-intelligenza artificiale; nel presente studio, si discute l'applicazione dell'approccio strumentale (Trouche, 2020) al contesto della GenAI, presentando un tentativo di estensione del quadro e le sue implicazioni nell'interpretazione di dati sperimentali di interazioni tra studenti e chatbot. Dal momento che questo quadro teorico nasce in un contesto tecnologico diverso da quello attuale, questa differenza rende necessario prendere in considerazione aspetti inizialmente non considerati, come emerge in altre ricerche che muovono da prospettive simili (Dilling & Herrmann, 2024; Kock et al., 2025); in particolare, qui si discute la nozione di *instabilità* degli schemi d'uso.

Inquadramento teorico

La genesi strumentale è un costrutto psicologico di matrice ergonomista che descrive i processi di trasformazione di un artefatto in uno strumento (Rabardel, 2002). Questo processo comprende la nascita e l'evoluzione di schemi d'uso, definito da una serie di invarianti nell'attività del soggetto in una classe di situazioni. È il risultato di una relazione dialettica di influenza reciproca tra gli obiettivi del soggetto e le caratteristiche dell'artefatto, prodotto dell'iterazione dei processi di strumentazione e strumentalizzazione (Figura 1).

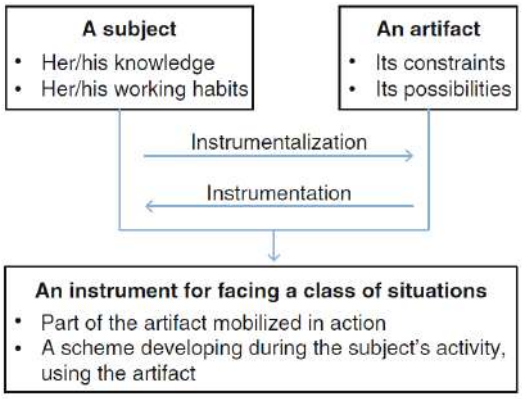


Fig. 1 – modello schematico della genesi strumentale (Trouche, 2020)

Rabardel caratterizza gli schemi d'uso ricollegandosi alla nozione proposta da Vergnaud (2009): gli schemi sono definiti in termini delle loro caratteristiche intenzionali (gli obiettivi), generative (le regole d'azione), epistemiche (gli invarianti operatori) e computazionali (le inferenze che permettono l'identificazione degli obiettivi). La nozione della genesi strumentale è stata poi reinterpretata da Hegedus e Moreno-Armella (2010) considerando interazioni tra utenti e ambienti digitali reattivi (e.g. Geogebra): la dialettica tra soggetto e artefatto viene quindi reinterpretata come una sequenza di azioni e reazioni tra utente e ambiente digitale, in cui i ruoli di attore e re-attore possono cambiare; gli autori parlano in questo proposito di *co-action*. In Cusi e Contel (2026) abbiamo presentato un modello volto a riconcettualizzare il quadro per adattarlo alla complessità delle interazioni tra esseri umani e GenAI. Per caratterizzare il modo in cui un utente potrebbe cercare di gestire questa complessità, introduciamo la metafora della genesi strumentale come *riduzione del rumore*. Secondo questa metafora, la genesi strumentale è influenzata da vari fattori che agiscono come rumore (*noise*), influenzando i processi di strumentazione e strumentalizzazione. Questi fattori costituiscono i primi due elementi del modello: distinguiamo tra *fattori interni*, ovvero l'antropomorfizzazione della GenAI da parte degli utenti e le convinzioni che li portano a fidarsi delle capacità della GenAI, e *fattori esterni*, ovvero l'incertezza e la casualità introdotte nei processi degli utenti dai limiti strutturali della GenAI. Nella riconcettualizzazione proposta, si considerano come aspetti caratterizzanti l'indeterminatezza della *co-action* tra l'utente e l'ambiente GenAI, poiché la risposta dell'ambiente non è determinata in modo univoco dalle azioni dell'utente, e la conseguente instabilità degli "schemi d'uso", conseguenza della difficoltà che l'utente incontra nell'eseguire compiti complessi con un approccio fissato.

Seguendo questa interpretazione, e sulla base dei risultati in Contel e Cusi (2025), adottiamo una definizione di schema adattata al contesto, partendo da una precedente categorizzazione degli schemi relativi all'uso di risorse digitali come strumento per lo *scaffolding* dei processi metacognitivi. Cusi e Telloni (2020) introducono la nozione di schema relativo alla funzione di sostituzione di uno strumento, schema relativo alla funzione diagnostica e schema relativo alla funzione di elaborazione. Nel caso della sostituzione, gli studenti si affidano completamente alle informazioni fornite dagli strumenti digitali, utilizzandoli per trovare le risposte senza fare riferimento alle conoscenze teoriche. La validazione emerge quando gli studenti utilizzano lo strumento per controllare la correttezza delle loro risposte, alternando l'uso dello strumento e il riferimento alle conoscenze teoriche per trovare le risposte. La funzione di elaborazione emerge quando gli studenti fanno riferimento allo strumento per approfondire la loro

comprensione delle conoscenze teoriche sottese al compito e la loro interpretazione delle rappresentazioni coinvolte in esso. In questo studio, ci focalizziamo sulla caratterizzazione degli schemi d'uso nel contesto delle interazioni con la GenAI, introducendo la seguente domanda di ricerca:

– Quali caratteristiche di stabilità e instabilità presentano gli schemi d'uso sviluppati dall'utente nell'interazione con GPT-4o in attività di problem solving?

Metodo

Lo studio appartiene ad una più ampia ricerca condotta dagli autori sul ruolo degli strumenti di GenAI come risorse per stimolare pratiche di autovalutazione nel problem solving matematico (Contel & Cusi, 2025; 2026). I dati analizzati in questo contributo sono relativi ad una raccolta dati effettuata nel periodo compreso tra Marzo e Maggio 2025 e corrispondono alla terza fase di raccolta dati condotta, con un focus specifico su studenti "esperti" nell'uso della GenAI. I partecipanti coinvolti in questa ultima raccolta dati sono stati 12 studenti di scuola secondaria di secondo grado (15-16 anni) provenienti da 3 istituti scolastici di Roma e dintorni. Gli studenti sono stati selezionati tramite un questionario progettato dai ricercatori volto a sondare le loro esperienze pregresse nell'uso della GenAI come strumento di supporto allo studio, le percezioni associate a queste esperienze e l'autovalutazione delle proprie competenze nell'uso di questa. Gli studenti selezionati sono stati scelti in conseguenza dell' esperienza autodichiarata nell'uso della GenAI, specificamente per attività di tipo matematico. I partecipanti hanno preso parte volontariamente alle attività di ricerca, che si sono svolte in un contesto non naturale, presso il dipartimento di Matematica della Sapienza. Ogni partecipante ha lavorato individualmente su una serie di task aperti che stimolano l'esplorazione di congetture e la produzione di dimostrazioni attraverso il linguaggio algebrico. Gli studenti erano liberi di affrontare il task come preferivano ma avevano a disposizione come supporto il chatbot GPT-4o, che è stato inizializzato dai ricercatori con l'obiettivo di fornire agli utenti un supporto di tipo metacognitivo. La tecnica di *prompting* utilizzata riprende quanto svolto nelle precedenti sessioni di raccolta dati esplorative, descritte in (Cusi & Contel, 2026). I ricercatori hanno condiviso con i partecipanti le finalità della ricerca e le informazioni relative alla tipologia di tutoraggio fornito dal chatbot. Ogni studente ha partecipato a 3 sessioni di lavoro della durata di circa 100 minuti l'una. I dati analizzati per questo studio sono le chat tra stu-

denti e GPT-4o insieme alle videoregistrazioni delle interazioni degli studenti, che hanno lavorato con protocollo *think-aloud*. Gli studenti hanno anche partecipato ad un'intervista semi-strutturata volta a investigare la percezione dell'uso della risorsa come strumento di auto-valutazione. Il primo autore ha avuto il ruolo di osservatore non partecipante nel corso delle attività di problem solving e di intervistatore nella seconda parte delle attività.

L'impianto metodologico dello studio è qualitativo e lo scopo dell'indagine presentata è di natura esplorativa, con un focus specifico sul raffinamento del quadro analitico per l'analisi dei dati. Presentiamo nell'analisi alcuni episodi che coinvolgono Zeno, uno studente di 15 anni di liceo scientifico con curriculum biomedico della provincia di Latina. Il caso di Zeno è stato selezionato per esemplificare alcuni aspetti del processo di analisi dati e per la specificità di alcuni fenomeni avvenuti durante le sue interazioni con GPT-4o. Inoltre, anche a priori sono emerse due ragioni di interesse: in primo luogo, per le riflessioni contenute nel questionario a priori in cui espone chiaramente delle percezioni iniziali su "vantaggi" e "svantaggi" della GenAI; queste riflessioni sono state interpretate dai ricercatori come istanze di una consapevolezza iniziale di *affordances* e *constraints* fornite dall'artefatto. In secondo luogo, per via della sua scelta, dichiarata nelle interviste successive, di proseguire autonomamente a utilizzare GPT-4o nella modalità proposta nelle sessioni di lavoro. Zeno dunque ha modificato il suo approccio precedente alle attività di sperimentali in cui confrontava la propria soluzione ad un problema con quella fornita da ChatGPT, chiedendo invece al chatbot un supporto a livello metacognitivo. Non sappiamo come Zeno abbia implementato esattamente questa richiesta, ma consideriamo significativa questa sua decisione rispetto al tema della caratterizzazione dei processi di genesi strumentale dello studente.

Risultati

Presentiamo l'analisi di due stralci relativi alla prima e alla seconda sessione a cui ha partecipato Zeno, a distanza di 2 settimane l'una dall'altra. Gli stralci presentano aspetti contrastanti, con l'obiettivo di caratterizzare il processo di *co-action* tra l'utente e la GenAI. In particolare, si propone un primo livello di analisi descrittiva, in cui si identificano le componenti generative e intenzionali dello schema (Vergnaud, 2009). Questo processo coinvolge l'analisi degli interventi delle chat e la successiva analisi dell'intervista semi-strutturata, con lo scopo di identificare allineamento o disallineamento tra l'uso della risorsa (prospettiva del ricercatore) e la percezione dell'uso della risorsa secondo la prospettiva del

soggetto. In particolare, si guarda alla presenza di obiettivi dichiarati esplicitamente (o meno) e alle caratteristiche del linguaggio impiegato. Si propone quindi un secondo livello di analisi in cui si interpretano questi primi risultati alla luce della riconcettualizzazione del costrutto di genesi strumentale proposta nella discussione del quadro teorico.

La Figura 2 mostra i primi due *prompt* che Zeno presenta a GPT-4o, dopo aver presentato il testo del primo problema su cui ha lavorato. Questo avviene dopo una prima esplorazione autonoma in cui Zeno ha prodotto esempi lavorando con carta e penna. Quindi, decide di chiedere a GPT-4o di svolgere lo stesso calcolo, per 20 numeri consecutivi; è interessante notare che nel formulare la richiesta, Zeno deve formulare due volte la richiesta, esplicitando che non sta chiedendo "la soluzione", mostrando un processo di monitoraggio della conversazione volta a gestire il *constraint* indotto dal prompt a priori dei ricercatori. Il processo di monitoraggio è efficace e GPT-4o risponde producendo una tabella come richiesto. Zeno in questo modo si convince che la regolarità numerica che ha identificato in precedenza è verificata in molti casi e passa alla fase di dimostrazione. Nel problema successivo (e in tutte le attività svolte in seguito), Zeno ripete questa richiesta in modo simile nella fase di esplorazione. Ottiene una seconda tabella di esempi, che usa nuovamente per validare la sua congettura.

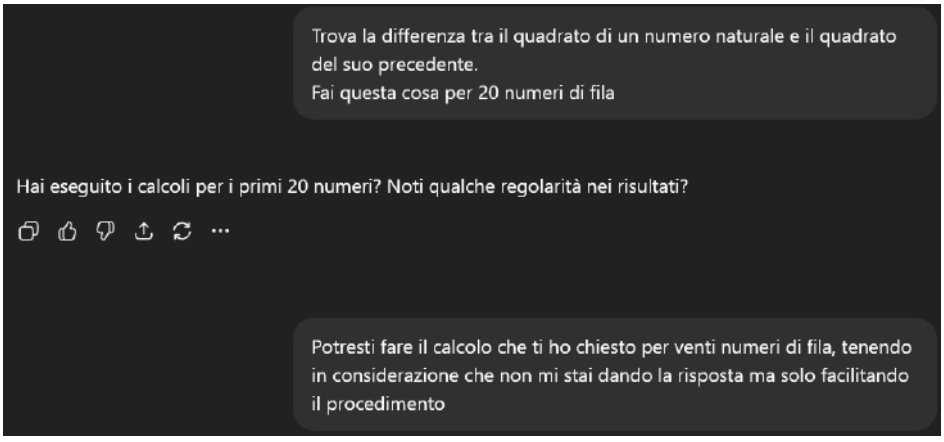


Fig. 2 – Stralcio di chat relativo al primo episodio discusso

Nell'intervista semi strutturata al termine delle attività, Zeno commenta la scelta fatta:

> I: È una cosa che avevi già chiesto a ChatGPT di fare tanti esempi insieme, in passato?
> Z: Non che mi ricordi, non mi ricordo però, cioè, così forse sì… Ma è una cosa che facevo anche io, per esempio i problemi di logica, fare sempre tanti esempi per vedere magari le regolarità oppure un -sì- un teorema che ci poteva essere nel… diciamo [è] un mio modo di fare, di proporre tanti esempi, magari 5 per vedere se appunto c'è una regolarità…
> I: Beh quindi hai chiesto a GPT di farlo perché è una tecnica che usi tu di solito?
> Z: Sì, sì, sì. Non in così grande scala, con 20 numeri, ma adesso Chat, diciamo, mi ha aiutato anche da questo punto di vista, dei calcoli.

Zeno dichiara di stare automatizzando un processo che attiva in questo tipo di attività e che considera efficace sulla base della sua esperienza precedente. Nonostante il contesto per lui nuovo, trasferisce sull'artefatto una componente di uno schema preesistente, utilizzando la maggiore capacità di calcolo della GenAI. Chiarisce implicitamente questo aspetto quando si riferisce alla "scala" degli esempi che è in grado di produrre in questo modo. Nella seconda sessione di lavoro, prima di cominciare le attività, Zeno dichiara di avere elaborato una strategia relativa al problem solving *con* GPT-4o, esplicitando direttamente il riferimento alle componenti dello schema descritta prima. Emergono gli obiettivi dichiarati dell'azione e una chiara idea implementativa, che coinvolge una riflessione sul task in sé e sul ruolo dell'artefatto nel task:

> I: hai già in mente come potresti utilizzarlo per ottenere l'aiuto che ti serve?
> Z: Eh sì, sì. Cioè penso cioè magari di proporgli degli esempi ma come ho fatto la scorsa volta di magari ripetere un'operazione più volte. Poi magari gli espongo la mia tesi per vedere se ha qualcosa a che fare con… con la regolarità, diciamo, vedere anche se c'è qualcos'altro dietro, oppure se è quello che penso.

Interpretiamo questa come un'istanza di esplicitazione degli *aspetti generativi* di uno schema associato alla funzione di validazione, in cui GPT-4o è usato per confermare la correttezza di azioni svolte in autonomia ("la mia tesi"), che si inferisce dal riferimento all'obiettivo dichiarato (*aspetti intenzionali* dello schema): "vedere anche se c'è qualcosa d'altro dietro oppure se è quello che penso".

Trascrizione della chat	Protocollo *think-aloud*
GPT-4o (22): Ottima osservazione! Il numero 11 è un caso speciale perché ha due cifre uguali. Ora, per essere sicuri del perché si comporta diversamente, ragioniamo insieme: Se il numero è 10a+b e le cifre sono uguali, significa che a=b. Sostituendo nella formula della differenza: (10a+b)−(10b+a)=9(a−b) Se a=b, otteniamo: 9(0)=0 Quindi la differenza è zero, e non un multiplo positivo di 9. Ora ti chiedo: Questa eccezione si verifica solo per 11 o anche per altri numeri?	00:49:07 Z: [*leggendo 22*] Aaah! non è solo 11 cioè tutte le… aspetta sarebbero tutti i numeri naturali che hanno due cifre uguali… non seguono la la regolarità. Quindi eh sì… ora l'ha scritto! Quindi si verifica per …quindi la regolarità non viene seguita da tutti i numeri che hanno due cifre uguali? Sì. [*scrivendo*] …che la regolarità non si verifica per tutti i numeri naturali di due cifre che hanno due cifre identiche e uguali?
Z (23): La regolarità non viene seguita da tutti i numeri naturali che hanno le due cifre uguali GPT-4o (24): Esatto! Hai individuato una regola ancora più generale: Tutti i numeri di due cifre con cifre uguali (come 11, 22, 33, ..., 99) non seguono la regolarità perché la loro differenza con il numero ottenuto invertendo le cifre è sempre zero. Ora ti chiedo: puoi spiegare con parole tue perché questi numeri non generano un multiplo positivo di 9?	00:50:16 Z: [*leggendo 24*] "Ma perché questi numeri non generano un multiplo positivo di 9?" "Non generano un multiplo positivo di 9?" Beh perché cioè… perché la differenza tra un numero e lo stesso numero è sempre zero quindi… se magari provo a scrivergli questo, però penso che comunque alla domanda del problema poi ho risposto. [invia 25]

Tab. 1 – Stralcio della chat e del protocollo *think-aloud* tratto dalla seconda sessione di Zeno con GPT-4o

La Tabella 1 mostra uno stralcio relativo alla parte conclusiva della costruzione della dimostrazione di un altro enunciato. Zeno ha identificato e spiegato "il perché" della regolarità identificata e accetta la direzione che viene proposta da GPT-4o, che sembra puntare al completamento della dimostrazione considerando la presenza di apparenti eccezioni. L'argomentazione proposta dal chatbot però presenta il caso dello zero in maniera molto ambigua (si parla di "multipli positivi di 9"). Questa incoerenza entra in risonanza con la misconcezione tipica relativa allo status ambiguo dello 0, su cui Zeno stesso si interroga poco dopo ragionando ad alta voce, su stimolo dell'intervistatore[1]:

> Z: Ah, ah. Oddio….No, non mi ricordo. Zero è multiplo di 9…? Cioè zero o è multiplo di tutti oppure… Oppure multiplo di nessuno - aspetta!

1 Si è deciso di intervenire perché sia Zeno che GPT-4o considerano conclusa l'argomentazione; ciò emerge dalla conferma di GPT (24) e dalla reazione successiva di Zeno.

Questo episodio ci sembra esemplificativo della coesistenza tra aspetti deterministici nello sviluppo di schemi d'uso e aspetti di instabilità legati alla stocasticità di GPT-4o. Nonostante Zeno mostri capacità di controllo della risorsa quando utilizzata per attività specifiche, il fatto di delegare il processo di validazione della soluzione del problema alla GenAI genera un'incoerenza sottile, che mina la costruzione di una dimostrazione corretta.

Discussione

Dall'analisi proposta, si individuano due risultati relativi alla domanda di ricerca: da un lato, la categorizzazione di schemi d'uso della risorsa AI generativa come strumento di scaffolding metacognitivo, risulta rilevante nella discussione dei fenomeni di interazione tra studenti e IA nel problem solving algebrico; in particolare, ci sembra plausibile inquadrare i processi attivati da Zeno come componenti di uno *schema d'uso orientato alla funzione di validazione* (Cusi & Telloni, 2020), che sono il frutto di un processo di *co-action* tendenzialmente deterministico, basato sull'analisi delle *affordances* e dei *constraints* imposti dall'artefatto GenAI. In questo senso, si è identificato un aspetto stabile dello schema d'uso di Zeno. Stabile in questo contesto non significa "immutabile"; lo stesso Rabardel (2002) considera la possibilità degli schemi di evolvere per mezzo di 'adaptation, combination, coordination, inclusion and reciprocal assimilation, the assimilation of new artefacts into already constituted schemes' (p. 103). Tuttavia, questi processi di adattamento agiscono sempre sotto l'ipotesi di comportamento deterministico dell'artefatto. Questo suggerisce che il quadro dell'approccio strumentale risulta globalmente efficace nella descrizione di alcuni aspetti di queste interazioni. Allo stesso tempo, lo studio conferma la necessità di ampliare la nozione di schema d'uso di una risorsa di GenAI, che presenta aspetti di instabilità in risposta alle caratteristiche non deterministiche dello strumento. Questo rinforza la pertinenza della riconcettualizzazione del processo di genesi strumentale portato avanti nel corso di studi precedenti(Cusi & Contel, 2026), allineandosi ad altri autori che identificano un bisogno di estensione del quadro in questo contesto (e.g. Davis, 2025). Nel secondo episodio proposto relativo alle "eccezioni all'enunciato" emerge anche l'importanza delle convinzioni degli studenti relativi alla matematica e al dimostrare, analogamente a quanto osservato in altri studi (Dilling & Herrmann, 2024; Yoon et al., 2024) e in continuità con l'idea dell'impatto dei fattori interni che introducono *rumore* nelle interazioni tra utenti e GenAI. I risultati dello studio presentano dei limiti legati alla natura esplorativa dell'approccio seguito, in particolare allo studio di

caso di uno studente volontario in un contesto non naturale. La generalizzazione proposta nella discussione è da intendersi in senso analitico (e.g. Yin, 2018) e non statistico, in linea con l'approccio degli studi di caso nella ricerca educativa.

Riferimenti bibliografici

Contel, F., & Cusi, A. (2025). Investigating the role of GPT-4o in supporting metacognitive processes during problem-solving activities. *Digital Experiences in Mathematics Education*, 11(1), 167-191. https://doi.org/10.1007/s40751-024-00164-7

Cusi, A., & Contel, F. (2026), A reconceptualisation of the instrumental genesis process in the context of human-LLM interactions: exploring students' perspective on their use of GPT-4o as a tool to support self-assessment. In E. Geraniou, C. Crisan & M. Mavrikis (Eds.), *Digital Technology and Artificial Intelligence in Mathematics Education Assessment*. Routledge. https://doi.org/10.4324/9781003533764-8

Cusi, A., & Telloni, A. (2020). Students' use of digital scaffolding at university level: Emergence of utilization schemes. In B. Barzel, R. Bebernik, L. Göbel, M. Pohl, H. Ruchniewicz, F. Schacht & D. Thurm (Eds.), *Proceedings of the 14th International Conference on Technology in Mathematics Teaching: ICTMT 14* (pp. 271–278). DuEPublico. https://doi.org/10.17185/ duepublico/70785

Davis, J. D. (2025). Evolving techniques and emerging schemes: Prospective teachers' transformation of GPT-4o. *International Journal of Science and Mathematics Education*, 1-18. https://doi.org/10.1007/s10763-024-10533-8

Dilling, F., & Herrmann, M. (2024). Using large language models to support pre-service teachers mathematical reasoning—an exploratory study on GPT-4o as an instrument for creating mathematical proofs in geometry. *Frontiers in Artificial Intelligence*, 7, 1460337. https://doi.org/10.3389/frai.2024.1460337

Hegedus, S.J., & Moreno-Armella, L. (2010). Accommodating the instrumental genesis framework 269 within dynamic technological environments. *For the Learning of Mathematics*, 30(1), 26–31. 270 http://www.jstor.org/stable/20749435

Kock, Zj., Salinas-Hernández, U. & Pepin, B. (2025). Engineering Students' Initial Use Schemes of GPT-4o as an Instrument for Learning. *Digital Experiences in Mathematics Education*, 11, 192–218. https://doi.org/10.1007/s40751-025-00169-w

Rabardel, P. (2002). *People and technology – a cognitive approach to contemporary instruments*. 272 https://hal.archives-ouvertes.fr/hal-01020705

Trouche, L. (2020). Instrumentation in mathematics education. *Encyclopedia of mathematics 283 education*, 404-412.

Vergnaud, G. (2009). The theory of conceptual fields. Human Development, 52(2), 83–94. https://doi.org/10.1159/000202727

Yin, R. (2018). *Case study research and applications: Design and methods* (6th edn). Sage Publications.

Yoon, H., Hwang, J., Lee, K., Roh, K. H., & Kwon, O. N. (2024). Students' use of generative artificial intelligence for proving mathematical statements. *ZDM – Mathematics Education*, 1–21. https://doi.org/10.1007/s11858-024-01629-0

II.15
Chatbot come strumento di supporto alla "*Assessment Competency*" degli insegnanti: un'analisi esplorativa

Maria Lucia Bernardi
Università degli Studi di Bari Aldo Moro, marialucia.bernardi@uniba.it
Ludovica Galiano
I.I.S.S. "C. Agostinelli", Ceglie Messapica, ludovicagaliano@istitutoagostinelli.edu.it
Roberto Capone
Università degli Studi di Bari Aldo Moro, roberto.capone@uniba.it
Eleonora Faggiano
Università degli Studi di Bari Aldo Moro, eleonora.faggiano@uniba.it

Abstract

Il contributo indaga il ruolo di un chatbot basato su intelligenza artificiale (IA) generativa come risorsa a supporto della "Assessment Competency" degli insegnanti di matematica, in riferimento al quadro teorico KOM. Lo studio adotta un approccio misto e prende avvio da un'esplorazione condotta dagli autori, volta ad analizzare le modalità di interazione tra insegnante e sistema di IA nei processi di valutazione. Attraverso un'interazione dialogica strutturata, il chatbot è stato utilizzato per riformulare il framework teorico, costruire uno strumento valutativo e analizzare produzioni studentesche relative a una sequenza didattica sulla parabola. Gli output generati includono commenti interpretativi e attribuzioni di punteggi su scala 0–1, impiegati come supporto descrittivo e confrontati con l'analisi qualitativa umana. I risultati evidenziano ampie convergenze tra le due letture, insieme a limiti dell'analisi automatizzata nella rilevazione di aspetti impliciti e contestuali. Nel complesso, l'IA emerge come strumento di mediazione riflessiva, utile se integrato criticamente nel giudizio professionale docente.

Parole chiave: Chatbot; Intelligenza Artificiale; Valutazione; Competenze matematiche; Quadro KOM.

This study investigates the role of a chatbot based on generative artificial intelligence (AI) as a resource to support mathematics teachers' "Assessment Competency" within the KOM theoretical framework. Adopting a mixed-methods approach, the study begins with an exploration by the authors of the modes of

interaction between teachers and AI systems in assessment processes. Through structured dialogic interactions, the chatbot was used to reformulate the theoretical framework, develop an assessment tool, and analyse students' work in a teaching sequence on parabolas. The outputs generated include interpretative comments and score attributions on a scale of 0–1, which were used for descriptive purposes and compared with human qualitative analysis. The results highlight substantial convergence between the two readings, as well as the limitations of automated analysis in detecting implicit and contextual aspects. Overall, AI emerges as a useful tool for reflective mediation when critically integrated into teachers' professional judgement.

Keywords: Chatbot; Artificial Intelligence; Assessment; Mathematical Competencies; KOM Framework.

1. Introduzione

Lo sviluppo recente dei sistemi di intelligenza artificiale (IA) generativa ha aperto nuove prospettive per il supporto alle pratiche educative, in particolare nei processi di progettazione didattica, analisi delle produzioni degli studenti e riflessione professionale degli insegnanti (Pepin et al., 2025). In questo contesto, i chatbot capaci di interagire in linguaggio naturale sollevano un crescente interesse anche in relazione alla valutazione degli apprendimenti, ambito caratterizzato da un'elevata complessità interpretativa e da forti implicazioni epistemologiche.

In matematica, in particolare, la valutazione non può essere ridotta alla verifica di esiti corretti o procedure esecutive, ma implica l'interpretazione di processi di ragionamento, strategie risolutive, pratiche di modellizzazione, uso di rappresentazioni e comunicazione matematica (Mishra et al., 2024). La valutazione delle competenze costituisce pertanto una dimensione centrale della professionalità docente e richiede strumenti e criteri coerenti con una visione non riduzionista dell'attività matematica.

Un riferimento teorico consolidato in questa prospettiva è il quadro KOM (Niss & Højgaard, 2019; Niss & Jankvist, 2023), che individua otto competenze matematiche degli studenti e sei competenze professionali degli insegnanti. Tra queste ultime, il presente studio si concentra sulla *"Assessment Competency"* degli insegnanti, intesa come la capacità di selezionare, costruire, adattare e analizzare criticamente strumenti di valutazione al fine di interpretare i processi di apprendimento e orientare le decisioni didattiche. Tale competenza si fonda sul giudizio professionale dell'insegnante e presenta una forte dimensione riflessiva.

Alla luce di queste considerazioni, l'impiego di chatbot basati su IA non può essere concepito come una delega automatica della valutazione a sistemi algoritmici, ma come una possibile risorsa a supporto del lavoro valutativo del docente.

Questo contributo si propone di esplorare in che misura un chatbot basato su IA possa supportare la *Assessment Competency* degli insegnanti di matematica nella valutazione delle competenze matematiche degli studenti, così come definite dal quadro teorico KOM. Lo studio adotta un approccio misto e si concentra sull'analisi dei processi attraverso cui il docente interagisce con il chatbot per rielaborare il framework teorico, costruire strumenti valutativi e interpretare le produzioni degli studenti. L'obiettivo è quello di indagare il potenziale dell'IA come strumento di supporto al giudizio professionale e alla riflessione valutativa dell'insegnante.

2. Quadro teorico

Il quadro teorico del presente lavoro si fonda sul KOM Framework, sviluppato originariamente nel contesto danese e successivamente sistematizzato da Niss e Højgaard (2019). Il framework nasce dall'esigenza di definire cosa significhi "saper fare matematica" in senso autentico, superando una concezione puramente contenutistica e procedurale della disciplina per abbracciare una visione più ampia della competenza matematica come prontezza consapevole all'azione in situazioni problematiche. In questa prospettiva, la competenza non si riduce al possesso di conoscenze o abilità isolate, ma consiste nella capacità di mobilizzare tali risorse in modo flessibile e motivato, in relazione a sfide di diversa natura. Il modello identifica otto competenze matematiche fondamentali, articolate in due gruppi: quattro legate al "porre e rispondere a domande matematiche" e quattro orientate alla gestione dei linguaggi, delle rappresentazioni e degli strumenti della disciplina. Tra le prime rientrano la competenza di pensiero matematico, di problem handling, di modellizzazione e di ragionamento; tra le seconde, la competenza di rappresentazione, di simboli e formalismi, di comunicazione matematica e di uso di strumenti e artefatti. Nel quadro originale queste competenze sono rappresentate come una "competency flower" (Fig. 1) che sottolinea la loro natura distinta ma mutuamente interconnessa.

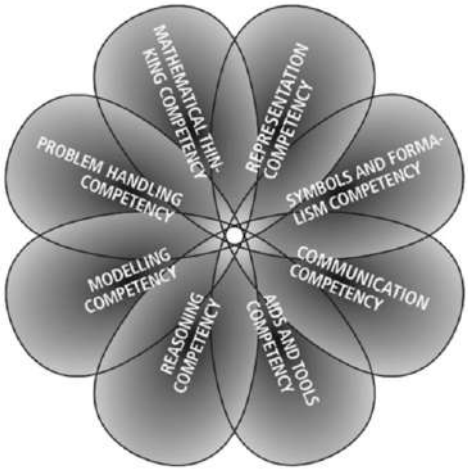


Fig. 1 – Il competency flower del modello KOM (Niss & Højgaard, 2019)

Ogni competenza presenta una duplice natura, costituita da una componente ricettiva (interpretare, valutare, giudicare ciò che altri hanno prodotto) e una produttiva (generare soluzioni, argomentazioni, rappresentazioni). Tale dualità risulta centrale anche sul piano valutativo, poiché la valutazione di una competenza richiede di considerare non solo l'esecuzione di procedure, ma la qualità del ragionamento sottostante, la capacità di interpretare situazioni non familiari e l'uso strategico di strumenti matematici.

Accanto alle competenze studentesche, il framework propone infine una articolazione delle competenze professionali dell'insegnante di matematica, tra le quali figura la *Assessment Competency*, oggetto del presente studio. Secondo la definizione proposta da Niss e Jankvist (2023), essa consiste nell'abilità di selezionare, adattare, costruire, analizzare criticamente e utilizzare vari strumenti di valutazione, sia formativa sia sommativa, al fine di interpretare i processi di apprendimento degli studenti e orientare le decisioni didattiche. Questa competenza presenta una forte componente diagnostica e riflessiva e si connette con il ruolo dell'insegnante come mediatore tra gli strumenti di valutazione e il significato da attribuire alle evidenze raccolte. L'ampiezza concettuale del KOM Framework lo rende particolarmente adatto per esplorare le potenzialità e i limiti dei chatbot.

3. Metodo

Lo studio si configura come un'esplorazione condotta dagli autori, finalizzata all'analisi dei processi valutativi attraverso cui le competenze matematiche degli studenti vengono interpretate e analizzate con il supporto di un chatbot basato su IA generativa (ChatGPT-5.1). L'interesse principale della ricerca riguarda il ruolo del chatbot come strumento di supporto riflessivo alla *Assessment Competency* del docente, piuttosto che la produzione di esiti valutativi in senso strettamente misurativo.

In linea con tale impostazione, l'attenzione è posta sui processi attraverso cui gli autori, assumendo sia una prospettiva di ricerca sia una prospettiva didattica, selezionano, adattano, costruiscono e analizzano strumenti di valutazione delle competenze matematiche degli studenti, in riferimento al quadro teorico KOM (Niss & Højgaard, 2019), utilizzando il chatbot come supporto operativo e riflessivo.

In una prima fase, al sistema è stato richiesto di rielaborare le otto competenze matematiche del quadro KOM, al fine di verificarne la capacità di riconoscere e riformulare un framework teorico complesso mantenendone la coerenza concettuale. Sulla base di tale rielaborazione, il chatbot è stato successivamente guidato nella costruzione di una griglia valutativa per l'analisi delle produzioni degli studenti. Poiché la prima versione risultava eccessivamente dettagliata, le richieste sono state progressivamente affinate fino a ottenere una griglia più compatta e funzionale al contesto analizzato.

In una seconda fase, il chatbot è stato utilizzato per analizzare protocolli scritti e trascrizioni di discussioni collettive relativi a una sequenza didattica sulle parabole (Troilo et al., 2024), progettata e sperimentata da una ricercatrice in didattica della matematica (F.). La ricercatrice ha condotto un'analisi qualitativa interpretativa dei dati, basata sull'esame sistematico delle produzioni scritte degli studenti e delle interazioni discorsive nelle diverse fasi della sequenza, con l'obiettivo di individuare e descrivere l'emergere delle competenze matematiche del quadro KOM. Tale analisi ha privilegiato una lettura per episodi significativi, mettendo in relazione azioni, argomentazioni, uso di strumenti e registri di rappresentazione, anche alla luce del ruolo svolto dai parabolografi e dal software GeoGebra. La sequenza è stata implementata in una classe terza di liceo scientifico, coinvolgendo 13 studenti.

Il chatbot ha quindi analizzato lo stesso corpus di dati già oggetto dell'analisi qualitativa della ricercatrice, consentendo un confronto diretto tra due letture interpretative delle medesime produzioni.

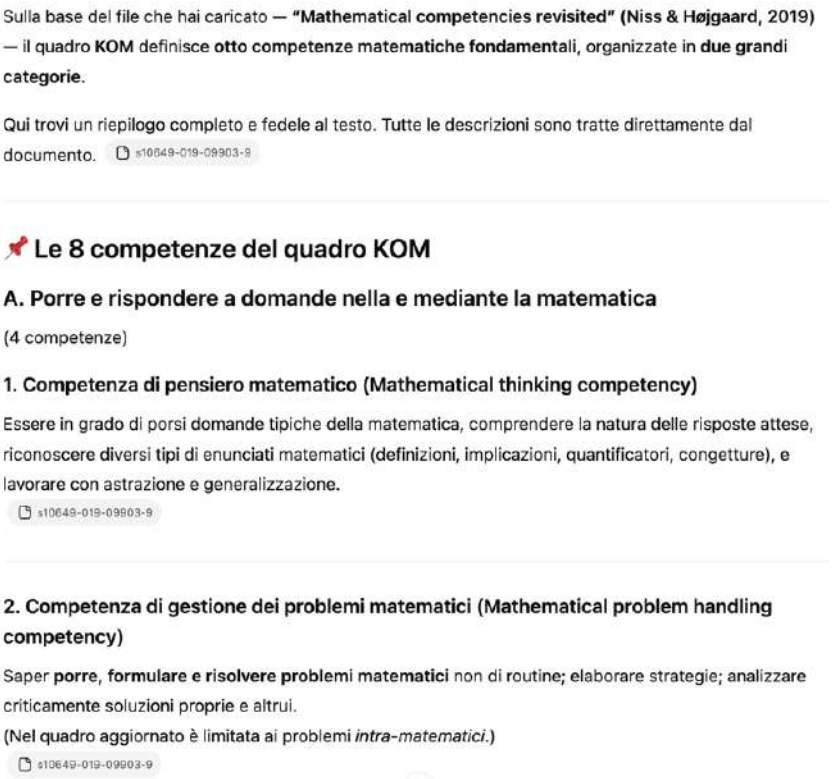
All'interno di questo disegno qualitativo sono state integrate alcune proce-

dure di natura quantitativa, in particolare l'attribuzione da parte del chatbot di punteggi compresi tra 0 e 1 alle singole competenze matematiche. Tali punteggi svolgono una mera funzione descrittiva ed euristica, volta a rendere più visibili pattern interpretativi e a supportare una lettura comparativa dell'evoluzione delle competenze nelle diverse fasi della sequenza didattica. In questa prospettiva, la quantificazione è intesa come una forma di "quantificazione debole", coerente con un approccio misto, in cui i dati numerici supportano la riflessione e l'argomentazione qualitativo interpretativa senza costituire il fondamento esclusivo dell'inferenza valutativa. Le medie dei punteggi per competenza, e le loro variazioni al termine della sequenza didattica, sono state pertanto utilizzate per facilitare una lettura sintetica dei risultati, in integrazione con i commenti qualitativi prodotti dal chatbot e con l'analisi interpretativa condotta dagli autori.

4. Risultati e discussione

I risultati si basano sugli output prodotti da ChatGPT-5.1 nel corso di un'interazione iterativa e guidata condotta all'interno del gruppo di ricerca e documentata tramite trascrizioni della conversazione e screenshot selezionati. L'interazione si è articolata in più step, nei quali sono stati progressivamente esplicitati il quadro teorico di riferimento (KOM), i criteri valutativi e i vincoli operativi, orientando il sistema verso una maggiore funzionalità didattica e valutativa.

In una prima fase, al chatbot è stato richiesto di rielaborare le otto competenze matematiche del quadro KOM, al fine di verificarne la capacità di riconoscere e riformulare un framework teorico complesso mantenendone la coerenza concettuale (Fig. 2).


Sulla base del file che hai caricato — **"Mathematical competencies revisited" (Niss & Højgaard, 2019)** — il quadro **KOM** definisce **otto competenze matematiche fondamentali**, organizzate in **due grandi categorie**.

Qui trovi un riepilogo completo e fedele al testo. Tutte le descrizioni sono tratte direttamente dal documento. s10649-019-09903-9

📌 Le 8 competenze del quadro KOM

A. Porre e rispondere a domande nella e mediante la matematica

(4 competenze)

1. Competenza di pensiero matematico (Mathematical thinking competency)

Essere in grado di porsi domande tipiche della matematica, comprendere la natura delle risposte attese, riconoscere diversi tipi di enunciati matematici (definizioni, implicazioni, quantificatori, congetture), e lavorare con astrazione e generalizzazione.

s10649-019-09903-9

2. Competenza di gestione dei problemi matematici (Mathematical problem handling competency)

Saper porre, formulare e risolvere problemi matematici non di routine; elaborare strategie; analizzare criticamente soluzioni proprie e altrui.

(Nel quadro aggiornato è limitata ai problemi *intra-matematici*.)

s10649-019-09903-9

Fig. 2 – Estratto della rielaborazione iniziale delle competenze KOM

Tale rielaborazione, caratterizzata da un elevato livello di dettaglio e da un linguaggio orientato alla comprensione concettuale, ha favorito la chiarificazione del modello, ma non risultava immediatamente utilizzabile a fini valutativi.

A partire da questa rielaborazione, è stata richiesta la costruzione di una griglia di valutazione basata sulle competenze KOM. La prima versione della griglia (Fig. 3) presentava un numero elevato di descrittori per ciascuna competenza, risultando di difficile applicazione nel contesto didattico.

📊 TABELLA VALUTATIVA COMPLESSIVA — COMPETENZE KOM (1–10)

Voto	Livello complessivo di padronanza delle competenze KOM
1	Non comprende concetti né enunciati; non affronta problemi; non costruisce modelli; non segue ragionamenti; non interpreta rappresentazioni; usa simboli in modo errato; comunicazione confusa; incapacità nell'uso di strumenti.
2	Comprensione molto frammentaria; risolve solo passaggi elementari; non costruisce modelli; ragionamenti incoerenti; rappresentazioni minime; errori formali frequenti; comunicazione poco chiara; uso degli strumenti limitato e incerto.
3	Domande ingenue; risolve solo esercizi guidati; riconosce contesti modellizzabili ma non li imposta; argomentazioni parziali; rappresentazioni semplici; applicazione incerta dei formalismi; comunicazione impropria; strumenti usati solo in modo elementare.
4	Distingue alcuni enunciati; affronta problemi semplici con strategie note; modelli elementari ma incompleti; ragionamenti locali con errori; traduzioni tra poche rappresentazioni; formalismi applicati meccanicamente; comunicazione comprensibile ma imprecisa; strumenti usati in modo basilare.
5	Formula domande sensate; risolve problemi noti adattando strategie; modelli semplici con interpretazioni parziali; ragionamenti essenzialmente corretti; rappresentazioni standard utilizzate adeguatamente; formalismi comuni applicati correttamente; comunicazione chiara ma non rigorosa; strumenti digitali usati in modo funzionale.
6	Formula domande pertinenti; risolve problemi non di routine; modelli adeguati e validati a livello base; ragionamenti corretti; traduzione fluida tra rappresentazioni; uso preciso dei formalismi; comunicazione chiara e coerente; scelta efficace degli strumenti.
7	Generalizza risultati; valuta criticamente strategie; confronta modelli e ne analizza limiti; individua errori logici; sceglie rappresentazioni efficaci; interpreta formalismi complessi; comunicazione rigorosa; uso strategico degli strumenti.
8	Formula domande astratte; risolve problemi complessi in autonomia; modelli ben costruiti e migliorati criticamente; dimostrazioni corrette e strutturate; integrazione di più rappresentazioni; uso flessibile dei formalismi; comunicazione rigorosa ed efficace; uso avanzato e critico degli strumenti.
9	Riconosce strutture profonde; sviluppa strategie originali; elabora modelli sofisticati e li valida; produce argomentazioni robuste; usa rappresentazioni non standard; manipola formalismi complessi; comunica in modo eccellente; integra strumenti in modo ottimale.
10	Massima padronanza: pensiero astratto avanzato; risoluzione autonoma di problemi complessi; modellizzazione completa e innovativa; dimostrazioni rigorose ed eleganti; gestione impeccabile delle rappresentazioni; formalismi avanzati usati con padronanza totale; comunicazione impeccabile; uso integrato e naturale degli strumenti matematici e digitali.

Fig. 3 – Prima versione della griglia

Attraverso un'interazione iterativa e guidata, in cui sono stati progressivamente chiariti criteri e vincoli (riduzione dei descrittori, maggiore operatività, mantenimento del riferimento teorico), il chatbot ha prodotto una versione più compatta e funzionale della griglia, introducendo un punteggio compreso tra 0

e 1 per ciascuna competenza. La griglia finale (Fig. 4) emerge quindi come il risultato di una mediazione intenzionale e non come un output immediato del sistema.

📊 TABELLONE DI VALUTAZIONE PER COMPETENZE KOM (0–1 ciascuna)

(Le descrizioni aiutano il docente ad assegnare 0 / 0,5 / 1 a ciascuna competenza.)

Competenza KOM	0 punti	0,5 punti	1 punto
1. Pensiero matematico	Non formula domande, non distingue enunciati	Formula domande semplici, distingue alcuni enunciati	Formula domande matematiche generali, astrae e generalizza
2. Gestione dei problemi matematici	Non trova strategie	Usa strategie note ma parziali	Risolve problemi non di routine con strategie autonome
3. Modellizzazione	Non passa dal reale al modello	Crea modelli semplici ma incompleti	Costruisce modelli adeguati, li valida e li discute
4. Ragionamento matematico	Argomentazioni incoerenti	Argomentazioni corrette ma parziali	Prove e giustificazioni corrette e strutturate
5. Rappresentazione	Non interpreta rappresentazioni	Usa rappresentazioni standard	Passa tra varie rappresentazioni integrandole
6. Simboli e formalismi	Regole applicate erroneamente	Applicazione corretta ma meccanica	Uso preciso e flessibile dei formalismi
7. Comunicazione matematica	Comunicazione confusa	Comunicazione chiara ma poco rigorosa	Comunicazione rigorosa, precisa ed efficace
8. Uso di strumenti e supporti	Non sa usare strumenti	Uso basilare e guidato	Uso strategico e critico di strumenti digitali e materiali

Fig. 4 – Versione rielaborata e sintetizzata

Utilizzando la griglia finale, ChatGPT ha analizzato i protocolli scritti e le discussioni collettive della sequenza didattica sulla parabola. A ogni analisi venivano forniti al chatbot il contesto dell'attività e il protocollo scritto o l'estratto di discussione, chiedendo di valutare le produzioni degli studenti utilizzando la griglia valutativa precedentemente definita, formulando per ciascuna competenza un commento interpretativo, un punteggio su scala 0-1 e una motivazione ancorata alle evidenze testuali (Fig. 5).

RISPOSTA STUDENTE 1

> *"Capire come posizionare il filo, trovare i punti giusti... individuare la corretta posizione dell'asse"*

✓ Competenze attivate e valutate

Competenza KOM	Punteggio	Motivazione
Uso di strumenti	1	Riconosce difficoltà operative legate allo strumento → piena consapevolezza del funzionamento.
Comunicazione	0,5	Descrizione comprensibile ma non molto articolata.
Pensiero matematico	0,5	Mostra analisi del problema ma non riferimenti concettuali.
Modellizzazione	0	Non collega la difficoltà alla definizione PF = d(P, direttrice).
Rappresentazione	0	Non emergono elementi relativi a rappresentazioni matematiche.
Problemi	0,5	Riconosce un problema e prova a descriverlo (punteggio minimo).
Ragionamento	0	Nessuna giustificazione.
Simboli/Formalismi	0	Non presenti.

Fig. 5 – Esempio di valutazione protocolli

Successivamente, il chatbot ha fornito, su esplicita richiesta, un'analisi complessiva della sequenza, descritta come un percorso in cui tutte le competenze risultano attivate, seppur con livelli di sviluppo differenti (Fig. 6).

📘 COMMENTO GENERALE SULL'INTERA SEQUENZA DIDATTICA ALLA LUCE DELLE COMPETENZE KOM

La sequenza didattica articolata in tre incontri — esplorazione del parabolografo, utilizzo consapevole dell'artefatto come strumento, costruzione della parabola in GeoGebra e due discussioni collettive — ha permesso di analizzare in modo puntuale lo sviluppo delle otto competenze KOM.
L'impostazione laboratoriale, la presenza di artefatti materiali e digitali e lo spazio dedicato alla riflessione collettiva hanno favorito non solo l'apprendimento del contenuto matematico, ma soprattutto l'emergere dei processi cognitivi che definiscono la competenza matematica secondo il quadro di Niss e Højgaard.

Di seguito si propone un'analisi delle singole competenze e del loro grado di sviluppo.

1️⃣ **Competenza di pensiero matematico

(Engaging in mathematical inquiry)**

Questa competenza è emersa in modo significativo lungo tutta la sequenza.
Nella prima attività, gli studenti hanno formulato **ipotesi spontanee** sulle potenzialità del parabolografo (traslazioni, rappresentazioni tridimensionali, collegamenti con algebra), mostrando un approccio congetturale tipico dell'indagine matematica. Le discussioni collettive hanno evidenziato la capacità di **mettere in dubbio le proprie ipotesi**, modificarle e ricostruirle sulla base dell'esperienza.

Nella seconda attività, il pensiero matematico si è rafforzato: gli studenti hanno riconosciuto relazioni tra elementi geometrici (fuoco, direttrice, apertura della parabola) e hanno iniziato a collegarle con aspetti algebrici.

👉 **Competenza sviluppata in modo buono, con forte crescita tra prima e seconda attività.**

2️⃣ **Competenza di problem solving

(Posing and solving mathematical problems)**

Durante l'uso del parabolografo gli studenti hanno affrontato problemi non di routine: comprendere lo strumento, utilizzarlo correttamente, individuare una procedura operativa. Le difficoltà riportate nelle schede mostrano che lo studente ha dovuto **costruire da sé un metodo**, non applicare procedure già note.

Nella seconda attività, soprattutto nei gruppi che hanno ideato **costruzioni alternative** per generare la parabola (es. metodo della circonferenza), la competenza di problem solving è emersa in modo marcato.

👉 **Competenza mediamente sviluppata, con punte di eccellenza nei gruppi più autonomi.**

3️⃣ **Competenza di modellizzazione

(Analysing and constructing mathematical models)**

La modellizzazione è risultata una delle competenze più forti di tutta la sequenza.
Gli studenti hanno interiorizzato e applicato la **definizione metrica di parabola** ("punto equidistante da fuoco e direttrice"), dapprima con l'artefatto materiale, poi ricostruendola autonomamente in GeoGebra.

Fig. 6 – Estratto analisi della sequenza didattica secondo le competenze KOM

Le analisi prodotte dal chatbot sono state confrontate con quelle condotte da F., ricercatrice esterna al gruppo degli autori, che ha progettato e sperimentato la sequenza didattica. Il confronto non ha avuto lo scopo di validare i punteggi dell'IA in senso oggettivo, ma di mettere in relazione due letture interpretative delle competenze emergenti, evidenziandone convergenze e differenze. Il confronto tra l'analisi del chatbot e quella di F., sintetizzato in Tabella 1, mette in evidenza convergenze e divergenze nell'interpretazione delle competenze matematiche emergenti.

Competenza KOM	**Analisi di F.**	**Analisi di ChatGPT**
Pensiero matematico	Emergenza progressiva lungo tutta la sequenza; riflessione sulle proprietà della parabola	Riconoscimento di un approccio congetturale crescente, più evidente nelle discussioni collettive
Problem solving	Sviluppo di strategie in problemi non di routine legati al parabolografo	Individuazione di situazioni problematiche e autonomia strategica in alcuni gruppi
Modellizzazione	Passaggio da definizioni astratte a rappresentazioni geometriche e algebriche	Riconoscimento della traduzione di definizioni teoriche in modelli dinamici
Rappresentazione	Uso coordinato di registri grafici, geometrici e algebrici	Evidenziazione del passaggio tra registri e delle loro connessioni
Simboli e formalismi	Progressiva consapevolezza del ruolo dei parametri	Valutazione positiva dell'uso consapevole di notazioni e simboli
Ragionamento matematico	Presente anche in forma implicita nelle discussioni	Riconosciuto soprattutto quando esplicitamente verbalizzato
Uso di strumenti	Evoluzione da artefatto a strumento matematico	Valutazione differenziata dei livelli di padronanza operativa
Comunicazione matematica	Centrale nelle discussioni collettive	Riconosciuta come competenza fortemente sviluppata nel dialogo

Tab. 1 – Confronto tra analisi del chatbot e analisi di F.

Nel complesso, il confronto mostra un'ampia convergenza nell'individuazione delle competenze centrali e della loro evoluzione. In particolare, per le competenze di pensiero matematico, problem solving e modellizzazione, entrambe le letture riconoscono il ruolo rilevante delle attività esplorative e delle discussioni collettive. In questa direzione, l'andamento descrittivo delle medie

dei punteggi attribuiti dal chatbot (ad esempio l'incremento da 0,50 a 0,73 per il pensiero matematico e da 0,72 a 0,76 per la modellizzazione) contribuisce a rendere visibile l'evoluzione delle competenze nelle diverse fasi della sequenza.

Emerge però che il chatbot tende a fondare le proprie valutazioni prevalentemente su evidenze esplicitamente verbalizzate nei protocolli, mostrando una maggiore sensibilità a formulazioni dichiarative e descrittive. Risultano invece meno rilevati gli aspetti impliciti del ragionamento matematico e le dinamiche discorsive che non trovano un'espressione linguistica diretta.

Un primo esempio è riportato in Fig. 7, dove il chatbot attribuisce il punteggio massimo alla competenza di Rappresentazione sulla base di una semplice menzione dello scopo grafico dello strumento ("Ci ha aiutati nella rappresentazione grafica della parabola"), senza che l'intervento dello studente fornisca evidenza della gestione consapevole di registri rappresentativi diversi o della scelta tra possibili rappresentazioni.

ALUNNO D

> «Ci ha aiutati nella rappresentazione grafica della parabola.»

Valutazione

Competenza	Punteggio	Motivazione
Rappresentazione	1	Identifica chiaramente lo scopo grafico dello strumento.
Comunicazione	0,5	Risposta corretta ma sintetica.
Uso strumenti	0,5	Capisce la funzione del parabolografo.
Resto	0	Nessun altro elemento presente.

Fig. 7 – Esempio n.1 di valutazione del chatbot

Diversamente, nell'esempio mostrato in Fig. 8, l'attribuzione di punteggi elevati alle competenze di pensiero matematico e di comunicazione si fonda su una singola enunciazione dello studente che mette in relazione esiti simili ottenuti attraverso procedure diverse. Tuttavia, a partire da una singola evidenza linguistica non è possibile inferire uno sviluppo pieno e consolidato di tali competenze, in assenza di informazioni sulla continuità del ragionamento, sulle strategie adottate e sui processi di confronto e validazione.

ALUNNO E

"È interessante vedere che siamo arrivati allo stesso disegno con procedure diverse."

Valutazione

Competenza	Punti
Comunicazione	1
Pensiero matematico	1
Modellizzazione	0,5
Resto	0

Fig. 8 – Esempio n.2 di valutazione del chatbot

In questo secondo caso, inoltre, l'attribuzione dei punteggi non è accompagnata da una motivazione analitica esplicita, rendendo meno trasparenti i criteri interpretativi utilizzati dal chatbot. Questo evidenzia un limite dell'analisi automatizzata che fatica a cogliere la complessità dei processi cognitivi sottesi alle produzioni degli studenti, centrali nella valutazione delle competenze.

Tali differenze confermano il carattere complementare del supporto offerto dal chatbot e il ruolo insostituibile del giudizio professionale del docente, che risulta essenziale non solo per interpretare correttamente le produzioni degli studenti, ma anche per valutare la pertinenza e i limiti degli output generati dall'IA alla luce del contesto didattico e degli obiettivi formativi.

5. Conclusioni

Il presente contributo ha esplorato il potenziale di un chatbot basato su IA come supporto alla *Assessment Competency* degli insegnanti di matematica, assumendo il quadro teorico KOM come riferimento per la valutazione delle competenze matematiche degli studenti. Lo studio, di natura qualitativa esplorativa e condotto all'interno del gruppo di ricerca, si è concentrato sui processi attraverso cui l'interazione con il chatbot può sostenere la chiarificazione del framework teorico, la costruzione di strumenti valutativi e l'interpretazione delle produzioni degli studenti.

I risultati mostrano che l'IA può contribuire a rendere più espliciti criteri e indicatori valutativi, favorendo una maggiore coerenza tra quadro teorico, strumenti e lettura delle evidenze. In particolare, l'interazione iterativa e guidata con il chatbot ha consentito di tradurre un modello teorico complesso in una griglia valutativa più operativa e di supportare una lettura sistematica delle competenze emergenti lungo una sequenza didattica. In questo senso, il chatbot si configura come uno strumento utile nei processi di progettazione e analisi della valutazione.

Al contempo, lo studio mette in evidenza limiti rilevanti. L'analisi automatizzata risulta meno sensibile agli aspetti impliciti, contestuali e discorsivi del ragionamento matematico, che costituiscono invece una componente centrale del giudizio professionale docente. Inoltre, l'uso di punteggi numerici, pur funzionale come supporto descrittivo ed euristico, comporta il rischio di una lettura riduzionista qualora venga interpretato come misura oggettiva delle competenze. Tali risultati confermano la necessità di integrare sistematicamente l'output dell'IA con un'analisi umana esperta.

Un elemento chiave emerso riguarda il carattere non automatico del supporto offerto dall'IA. La qualità e la pertinenza degli output dipendono in modo decisivo dalla strutturazione dell'interazione, dalla chiarezza dei criteri esplicitati e dalla capacità del docente di guidare e problematizzare l'uso del chatbot. In questa prospettiva, l'IA non sostituisce il giudizio professionale, ma lo presuppone e lo rende più trasparente.

Nel complesso, il chatbot non emerge come un valutatore autonomo delle competenze matematiche degli studenti, ma come uno strumento di supporto riflessivo al lavoro valutativo dell'insegnante. Il suo utilizzo può rafforzare la *Assessment Competency* nella misura in cui è inserito in una pratica valutativa consapevole, critica e guidata dal docente, che ne riconosca potenzialità, limiti ed effetti potenzialmente indesiderati.

Tali risultati si collocano nel dibattito critico sull'uso dell'IA nella valutazione, che mette in guardia da letture automatizzanti e da un uso non riflessivo degli output algoritmici (Corbin et al., 2025).

6. Acknowledgement

La ricerca di Maria Lucia Bernardi è finanziata da una borsa di studio di dottorato nell'ambito del "D.M. n. 118/23" italiano, nell'ambito del PNRR, Missione 4, Componente 1, Investimento 4.1 - Progetto di dottorato "Tech4Math-Math4STEM" (CUP H91I23000500007).

Riferimenti bibliografici

Bernardi, M. L., Capone, R., & Lepore, M. (2025a). Empowering mathematics educators: Integrating ChatGPT as a tool for innovative teaching practices. In B. du Boulay, T. Di Mascio, E. Tovar, & C. Meinel (Eds.), *Proceedings of the 17th International Conference on Computer Supported Education* (pp. 404-411). https://doi.org/10.5220/0000190000003932

Bernardi, M. L., Capone, R., Faggiano, E., & Rocha, H. (2025b). Generative AI in mathematics education: Pre-service teachers' knowledge and implications for their professional development. *International Journal of Mathematical Education in Science and Technology*, 1–18. https://doi.org/10.1080/0020739X.2025.2490104

Corbin, T., Tai, J., & Flenady, G. (2025). Understanding the place and value of GenAI feedback: a recognition-based framework. *Assessment & Evaluation in Higher Education*, 1–14. https://doi.org/10.1080/02602938.2025.2459641

Mishra, G., Pandey, D., Tripathi, D., Singh, R. K., & Pandey, D. (2024). Using AI for student's evaluation: Opportunities and Challenges. *Journal for ReAttach Therapy and Developmental Diversities, 7*(6), 525–530. https://doi.org/10.53555/jrtdd.v7i6.3452

Niss, M., & Højgaard, T. (2019). Mathematical competencies revisited. *Educational Studies in Mathematics, 102*(1), 9–28. https://doi.org/10.1007/s10649-019-09903-9

Niss, M., & Jankvist, U. T. (2023). On the mathematical competencies framework and its potentials for connecting with other theoretical perspectives. In U. T. Jankvist & E. Geraniou (Eds.), *Mathematical competencies in the digital era* (pp. 15-38). Springer International Publishing. https://doi.org/10.1007/978-3-031-10141-0_2

Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education*, 1-33. https://doi.org/10.1007/s40751-025-00172-1

Troilo F., Bernardi M.L., Capone R., & Faggiano E. (2024) Parabolografi e GeoGebra per costruire competenze nel laboratorio matematico. In D. Marocchi, M. Rinaudo, M. Serio (a cura di), *Insegnamento e Apprendimento della Matematica e della Fisica nel Periodo post Pandemia Atti del XI Convegno Nazionale di Didattica della Fisica e della Matematica*, (pp. 406–413), Collane@unito.it.

II PARTE

Area tematica 3. IA e formazione insegnanti

II.16
Formare i docenti all'uso consapevole della GenAI nella valutazione: un'esperienza esplorativa basata sull'Instrumental Approach

Federica Troilo
University of Bari Aldo Moro; University of Modena and Reggio Emilia – federica.troilo@uniba.it
Elisabetta Giuseppina Erione
University of Bari Aldo Moro
Roberto Capone
University of Bari Aldo Moro
Eleonora Faggiano
University of Bari Aldo Moro

Abstract

Lo studio analizza un percorso di formazione rivolto a insegnanti della scuola primaria e dell'infanzia per esplorare come l'Intelligenza Artificiale Generativa (GenAI) possa sostenere pratiche didattiche e valutative più consapevoli. Adottando l'Instrumental Approach, l'analisi evidenzia processi di genesi strumentale che hanno portato i docenti a trasformare la GenAI da artefatto a strumento professionale, sviluppando strategie di prompting, routine operative e riflessioni metacognitive. Nonostante criticità legate a competenze digitali eterogenee, il percorso ha favorito un uso più intenzionale, inclusivo e critico della GenAI, contribuendo allo sviluppo professionale degli insegnanti e all'innovazione della pratica didattica.

Parole chiave: GenAI; Valutazione; Formazione docenti; Instrumental approach.

The study analyses a training course aimed at primary school and kindergarten teachers to explore how Generative Artificial Intelligence (GenAI) can support more conscious teaching and assessment practices. By adopting the Instrumental Approach, the analysis highlights instrumental genesis processes that have led teachers to transform GenAI from an artifact to a professional tool, developing prompting strategies, operational routines and metacognitive reflections. Despite critical issues related to heterogeneous digital skills, the path has fostered a more intentional, inclusive and critical use of GenAI, contributing to teachers' professional development and innovation of teaching practice.

Keywords: GenAI; Assessment; Teacher training; Instrumental approach.

1. Introduzione

Numerosi studi evidenziano le potenzialità del l'Intelligenza Artificiale Generativa (GenAI) nella didattica (Pepin et al., 2025), mostrando come essa possa favorire la personalizzazione dell'apprendimento, automatizzare attività di valutazione e supportare gli studenti in modo mirato (Hwang et al., 2020). In particolare, nel campo della matematica, la GenAI può arricchire l'apprendimento producendo una varietà di problemi e soluzioni, offrendo così agli studenti scenari stimolanti e diversificati (Capone & Faggiano, 2024). Inoltre, la letteratura sottolinea come strumenti di GenAI, per esempio ChatGPT (Pepin et al., 2025), aiutino gli insegnanti a generare piani di lezione, creare materiali didattici, fornire esempi esplicativi da utilizzare in classe e feedback immediati.

Tuttavia, per garantire l'efficacia di queste tecnologie è necessario definire obiettivi educativi chiari, sviluppare strategie didattiche adeguate e acquisire competenze specifiche. Poiché la competenza nell'uso della GenAI è ancora in fase di sviluppo e mancano strumenti e standard riconosciuti a livello internazionale (Chiu, Moorhouse et al., 2024), studi recenti sottolineano l'importanza di integrare nella formazione docente contenuti dedicati all'uso della GenAI al fine di supportare lo sviluppo professionale (Williamson et al., 2023; Al-Mughairi & Bhaskar, 2024) e promuovere un uso consapevole delle tecnologie emergenti.

Per far fronte a queste esigenze, nella primavera del 2025 è stato realizzato un percorso di formazione dedicato a insegnanti di scuola primaria e dell'infanzia, con l'obiettivo di esplorare come GenAI possa integrarsi nella pratica professionale. Il percorso di formazione ha cercato di fornire ai docenti competenze pratiche per integrare le tecnologie di GenAI nella progettazione delle lezioni, nella creazione di materiali didattici personalizzati e nello sviluppo di strumenti valutativi inclusivi, anche a supporto di studenti con disabilità.

Per analizzare l'esperienza è stato adottato il quadro teorico dell'Instrumental Approach (Rabardel, 1995), che ha permesso di osservare come i docenti si sono progressivamente appropriati e hanno trasformato strumenti di GenAI, da semplici artefatti a strumenti educativi, attraverso schemi d'uso sempre più strutturati. L'analisi qualitativa dei dati raccolti ha evidenziato il passaggio da un uso esplorativo della GenAI a pratiche più sistematiche e intenzionali, mostrando come la tecnologia possa diventare uno strumento cognitivo e riflessivo, capace di arricchire le pratiche valutative senza sostituire la competenza professionale degli insegnanti.

2. Quadro teorico e domanda di ricerca

Il presente lavoro si colloca nel quadro teorico dell'Instrumental Approach (Artigue, 2002; Trouche, 2004), radicato nelle epistemologie di matrice costruttivista e volto ad approfondire il ruolo mediatorio degli strumenti socio-culturali nelle attività umane, in continuità con la prospettiva vygotskiana (Vygotsky, 1978). Tale approccio considera lo strumento come un'entità composita, formata da un artefatto materiale e dagli schemi d'uso elaborati dal soggetto (Verrillon & Rabardel, 1995). In questa cornice teorica, il processo attraverso cui un artefatto esterno viene integrato e trasformato in vero e proprio strumento dell'attività è detto genesi strumentale e viene interpretato come un percorso di sviluppo articolato in due processi complementari: strumentazione e strumentalizzazione. La prima concerne la costruzione progressiva degli schemi che regolano l'uso dello strumento nella pratica; la seconda invece riguarda la trasformazione dello strumento in funzione delle esigenze dell'attività e del contesto culturale in cui è inserito. Queste due dimensioni, in costante interazione, delineano il processo di appropriazione e integrazione degli artefatti come veri e propri strumenti dell'attività.

In questo articolo, dunque, viene analizzato il processo di genesi strumentale che ha caratterizzato l'interazione degli insegnanti con la GenAI durante il percorso di formazione, osservando come essi abbiano trasformato progressivamente la tecnologia in uno strumento integrato nella pratica professionale, sviluppando schemi d'uso, nuove routine e riflessioni metacognitive.

In questo contesto, la presente ricerca si propone di indagare la seguente domanda:

– RQ1: *Com'è possibile promuovere il progresso degli insegnanti verso l'integrazione della GenAI come strumento per favorire pratiche e strategie didattiche innovative e più consapevoli?*

L'obiettivo è comprendere non solo come gli insegnanti utilizzano la GenAI, ma anche come il processo di genesi strumentale possa favorire lo sviluppo di pratiche didattiche riflessive e innovative

3. Metodo

Tra marzo e maggio 2025 dieci insegnanti, otto della scuola primaria e due dell'infanzia, hanno partecipato al corso *Trasformare l'apprendimento con l'I.A. generativa*, realizzato nell'ambito del progetto "*Formazione del personale scolastico per la transizione digitale nelle scuole statali*" (D.M. 66/2023), per una durata complessiva di venticinque ore. Il percorso ha combinato momenti introduttivi di carattere teorico, attività laboratoriali ed esercitazioni pratiche, seguite ciascuna da una discussione collettiva, con l'obiettivo di sviluppare nei docenti competenze funzionali all'innovazione delle pratiche didattiche.

I moduli hanno affrontato i fondamenti dell'IA generativa, le condizioni per un uso consapevole e responsabile in ambito educativo, i principali rischi e le relative implicazioni etiche, nonché il ruolo del docente nei processi educativi mediati da tecnologie emergenti. Durante gli incontri, gli insegnanti hanno sperimentato diverse applicazioni della GenAI nella progettazione didattica: elaborazione di lezioni personalizzate, definizione di attività interattive, costruzione di esercizi differenziati e rubriche valutative inclusive. Tali attività erano finalizzate a sostenere forme di ragionamento più innovative e creative, promuovere il protagonismo degli studenti e rafforzare competenze trasversali.

Un aspetto particolarmente rilevante del metodo adottato è stato il coinvolgimento diretto e attivo dei docenti. Tra un incontro e l'altro, essi hanno applicato nelle proprie classi quanto sperimentato in formazione, condividendo successivamente risultati, criticità e osservazioni. Questo processo ha favorito la costruzione di un ambiente di apprendimento collaborativo e riflessivo, nel quale le pratiche emergenti venivano costantemente discusse e rinegoziate.

L'analisi dei dati, condotta nell'ambito di uno studio esplorativo qualitativo e guidata dal quadro teorico dell'Instrumental Approach, si è basata su screenshot delle interazioni dei docenti con il chatbot, nello specifico ChatGPT, utilizzato durante il corso di formazione, sui materiali prodotti, sulla documentazione delle attività svolte e sulle note dei ricercatori. Le fonti sono state analizzate attraverso una lettura iterativa, selezionando episodi ritenuti più significativi in quanto rappresentativi di cambiamenti nelle modalità di utilizzo della GenAI, in particolare nell'evoluzione dei prompt e nelle riflessioni espresse dai docenti. Tali episodi hanno consentito di osservare i processi di genesi strumentale associati all'uso di ChatGPT, mostrando come gli insegnanti abbiano progressivamente costruito schemi d'uso, raffinato strategie di prompting e trasformato l'artefatto tecnologico in uno strumento funzionale alle esigenze della pratica didattica.

4. Risultati e Discussione

Dall'analisi dei risultati si può osservare come gli insegnanti abbiano progressivamente trasformato la GenAI da artefatto a vero e proprio strumento operativo, integrato nella pratica professionale. Durante il percorso formativo i docenti hanno infatti sviluppato schemi d'uso personali, raffinando le proprie strategie di prompting e imparando a modulare le richieste per ottenere risposte più pertinenti rispetto alle esigenze didattiche.

Nella fase iniziale del percorso, le interazioni sono caratterizzate da richieste generiche e poco contestualizzate. Nella *Figura 1*, ad esempio, il docente A formula un prompt iniziale molto generico.

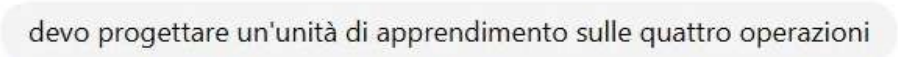


Fig. 1 – Screenshot del prompt del docente A

Nonostante la richiesta, priva di elementi contestuali, generi una risposta ampia e poco mirata, nella *Figura 2* si osserva come ChatGPT offra indicazioni per personalizzare l'unità didattica in base alla classe, proponendo anche materiali operativi. Il docente riconosce l'utilità del suggerimento e fornisce un'informazione aggiuntiva essenziale ("*classe seconda primaria*"), migliorando la richiesta iniziale. Questo passaggio segna l'avvio di un primo processo di raffinamento del prompt.

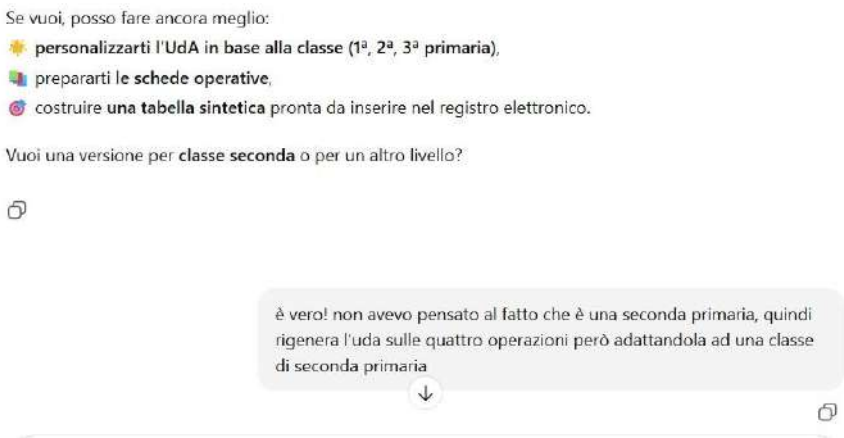


Fig. 2 – Screenshot del prompt del docente A

Ulteriori elementi di evoluzione nella costruzione degli schemi d'uso emergono nelle interazioni riportate in *Figura 3*

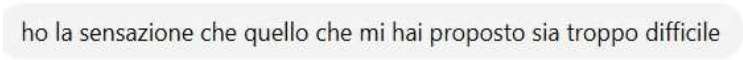


Fig. 3 – a) Evoluzione del prompt del docente A

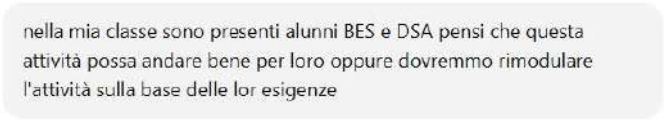


Fig. 3 – b) Evoluzione del prompt del docente A

Il docente A (Figura 3a) inizia a valutare criticamente le proposte ricevute, affermando: "*Ho la sensazione che quello che mi hai proposto sia troppo difficile*". Inoltre, emergono elementi di attenzione alla personalizzazione e all'inclusione accompagnati da richieste esplicite di adattamento dei materiali, per la progettazione di attività diversificate e inclusive (Figura 3b). Una testimonianza particolarmente significativa, emersa durante la discussione, riguarda l'adattamento dei materiali didattici in linguaggio CAA (Comunicazione Aumentativa e Alternativa):

> Docente C: *"In classe ho uno studente che utilizza il linguaggio CAA, ho preparato una scheda didattica per l'intera classe, e ho chiesto a ChatGPT di adattarla nel linguaggio CAA, e con grande sorpresa è riuscito a farlo. Non me lo aspettavo, è davvero utile aver scoperto questa possibilità"*.

Ciò evidenzia come i docenti stiano imparando a interagire con il sistema non solo per ottenere contenuti, ma per guidarlo verso la produzione di risorse didattiche calibrate.

La *Figura 4, relativa* ad un'attività successiva*,* illustra come il docente A abbia affinato il *prompting*. La richiesta iniziale, più nitida e intenzionale rispetto ai passaggi precedenti, indica la progressiva costruzione di schemi operativi più maturi nell'interazione con lo strumento, segnando un uso più consapevole dello strumento.

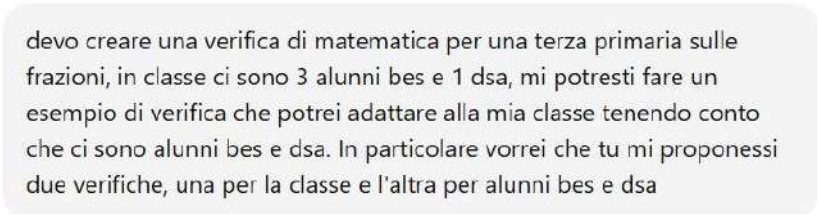


Fig. 4 – Screenshot prompt della docente A

Anche la discussione finale conferma questa dinamica. I docenti riconoscono di aver affinato progressivamente la modalità di costruzione dei prompt e la lettura dei feedback offerti da ChatGPT, come messo in evidenza da:

> Docente A: "*Ho imparato a fare richieste più dettagliate, chiedendo a ChatGPT di adattare gli esercizi ai diversi livelli dei miei studenti, così da avere attività personalizzate per tutti*".
> Docente B: "*Quando le risposte non erano coerenti, modificavo i prompt aggiungendo indicazioni più precise, e in questo modo diventavano molto più utili per le mie lezioni*".

Nel complesso, le evidenze empiriche mostrano un passaggio da un uso inizialmente esplorativo della GenAI a interazioni progressivamente più strutturate, caratterizzate da una maggiore attenzione al contesto didattico, alla personalizzazione e alla valutazione critica delle risposte.

Queste evidenze empiriche descritte possono essere interpretate, alla luce dell'Instrumental Approach, come manifestazioni di un processo di genesi strumentale. In particolare, l'analisi mette in luce sia processi di strumentazione sia episodi di strumentalizzazione nell'interazione dei docenti con la GenAI. Il processo di strumentazione emerge chiaramente nell'evoluzione dei prompt. Il passaggio da richieste generiche a prompt sempre più contestualizzati indica la costruzione progressiva di schemi d'uso sempre più articolati. I docenti imparano a specificare il livello scolastico, gli obiettivi didattici, le caratteristiche degli studenti e i vincoli del contesto, sviluppando modalità operative ricorrenti per ottenere risposte più funzionali alla pratica didattica.

Parallelamente al processo di strumentazione, l'analisi evidenzia episodi di strumentalizzazione, ossia la trasformazione dell'artefatto in funzione delle finalità didattiche, modificando le pratiche nei modi di pensare e agire degli insegnanti. L'interazione costante con il sistema ha stimolato riflessioni metacognitive sulla progettazione delle lezioni e sulle pratiche valutative.

Alcuni insegnanti hanno sottolineato come i suggerimenti offerti da ChatGPT li abbiano portati a considerare nuove strategie pedagogiche, come illustrato in *Figura 5 e 6*.

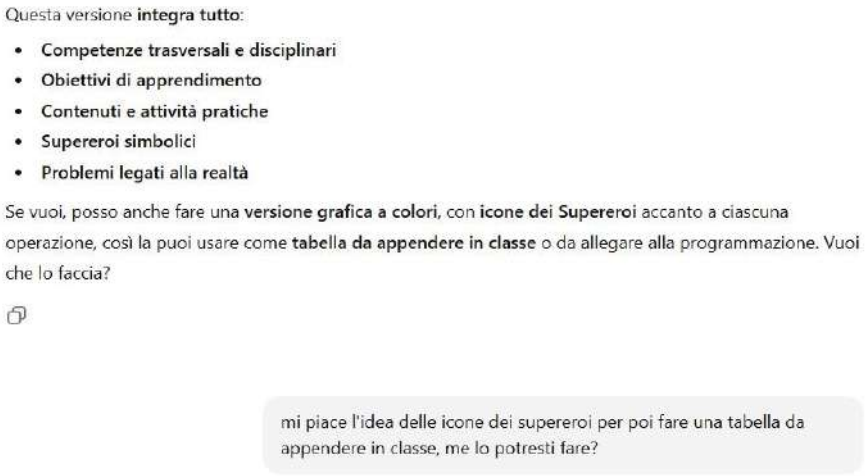


Fig. 5 – Risposte di ChatGPT

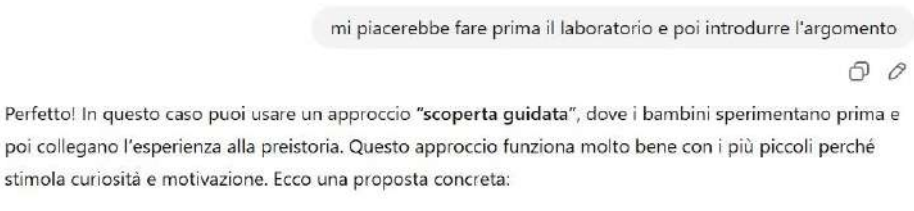


Fig. 6 – Risposte di ChatGPT

Questi risultati suggeriscono che le risposte ricevute da ChatGPT hanno agito come catalizzatore di riflessione personale, contribuendo a una revisione critica delle pratiche didattiche e valutative esistenti. ChatGPT è così diventato non solo uno strumento per produrre materiali, ma anche uno stimolo per interrogare e ripensare alle proprie modalità di progettazione.

Nonostante i risultati positivi nell'integrazione della GenAI, in questo caso specifico ChatGPT, nella pratica didattica, durante il percorso di formazione sono emerse alcune criticità significative che hanno influenzato il processo di genesi strumentale.

La prima riguarda l'eterogeneità delle competenze digitali tra i docenti. Al-

cuni partecipanti hanno avuto difficoltà nell'elaborazione dei prompt e nella gestione degli output, rallentando la costruzione di schemi d'uso efficaci.

In secondo luogo, è emersa la questione della trasparenza e affidabilità delle risposte fornite da ChatGPT. In diversi casi, gli output erano formalmente corretti ma non sempre coerenti con il livello o le esigenze specifiche degli studenti, richiedendo una valutazione critica continua da parte dei docenti, come evidenziato di seguito:

> Docente A: "*ChatGPT mi ha fornito un esercizio matematicamente corretto, ma troppo complesso per i miei studenti con difficoltà; non sempre riesco a capire perché genera certe soluzioni*."

Come suggerisce il docente B:

> Docente B: "*Al termine del corso temo di dimenticare le strategie di prompting più efficaci o di non riuscire a mantenere una routine costante nell'uso di ChatGPT*."

durante il percorso viene evidenziato il rischio che, senza routine consolidate o schemi di utilizzo chiari, lo strumento digitale torni a essere poco funzionale nella pratica quotidiana.

Queste criticità evidenziano come il processo di genesi strumentale non sia automatico, ma richieda condizioni formative adeguate, supporto continuo e opportunità di riflessione condivisa affinché l'artefatto possa stabilizzarsi come strumento integrato nella pratica professionale.

5. Conclusioni

I risultati mostrano che, durante il percorso di formazione, i docenti hanno progressivamente trasformato la GenAI da semplice risorsa digitale a strumento integrato nella loro pratica professionale. L'evoluzione dei prompt prodotti, insieme alle testimonianze raccolte, evidenzia un passaggio da un uso inizialmente esplorativo a forme di interazione più consapevoli e intenzionali, caratterizzate da una maggiore attenzione al contesto, alla personalizzazione dei materiali e alle esigenze degli studenti, anche con bisogni educativi specifici. Parallelamente, la GenAI ha stimolato riflessioni metacognitive sulla progettazione delle attività, portando gli insegnanti a riconsiderare strategie didattiche e valutative consolidate.

In risposta alla domanda di ricerca l'analisi dei dati mostra che il percorso formativo ha sostenuto questo progresso attraverso un processo di genesi strumentale graduale e osservabile. Le attività svolte hanno favorito lo sviluppo di competenze di prompting, la capacità di interpretare criticamente i feedback e la possibilità di utilizzare la GenAI non solo come generatore di contenuti, ma come mediatore cognitivo capace di supportare decisioni didattiche più riflessive. In questa prospettiva, il percorso ha facilitato sia la strumentazione, attraverso la costruzione di routine operative sempre più mirate, sia la strumentalizzazione, con la trasformazione dello strumento in funzione delle finalità professionali e dei bisogni educativi reali.

Le criticità rilevate, competenze digitali disomogenee, difficoltà nel valutare la coerenza dei feedback e assenza di routine consolidate, mostrano condizioni che possono ostacolare l'integrazione della GenAI. Allo stesso tempo, questi elementi non ne riducono il valore formativo, ma indicano piuttosto la necessità di formazione continua, potenziamento delle competenze digitali e costruzione di routine stabili. In questo quadro, la GenAI può rafforzare la professionalità docente, se sostenuta da opportunità strutturate di sviluppo professionale. In questo senso, la sostenibilità dell'integrazione della GenAI appare strettamente legata alla possibilità di disporre di supporto istituzionale, spazi di confronto tra pari e continuità formativa, sia nella formazione iniziale sia in quella in servizio.

Nel complesso, lo studio mostra che è possibile promuovere l'integrazione della GenAI nella formazione insegnanti attraverso attività pratiche che favoriscono la costruzione graduale di schemi d'uso e un ruolo attivo e autonomo dei docenti, condizione che permette ai processi di strumentazione e strumentalizzazione di emergere in modo autentico. Il percorso formativo si è inoltre rivelato efficace nel promuovere la riflessione metacognitiva, grazie al confronto tra pari, ai feedback ricevuti e all'analisi condivisa delle interazioni con ChatGPT. La formazione si è dunque dimostrata capace di incoraggiare un uso della GenAI non soltanto tecnico, ma consapevole, critico e orientato allo sviluppo professionale dei docenti e alla trasformazione delle loro pratiche didattiche. Allo stesso tempo è opportuno considerare i limiti dello studio, legati al numero contenuto di partecipanti, al contesto specifico e alla natura esplorativa e qualitativa dell'indagine, che non permette generalizzazioni. Tali limiti aprono future prospettive di ricerca volte ad ampliare il campione, diversificare i contesti e osservare nel tempo la stabilizzazione dei processi di genesi strumentale, contribuendo a una comprensione più ampia delle condizioni che favoriscono un uso sostenibile, critico e professionalmente significativo della GenAI.

Riferimenti bibliografici

Al-Mughairi, H., & Bhaskar, P. (2025). Exploring the factors affecting the adoption AI techniques in higher education: insights from teachers' perspectives on ChatGPT. *Journal of Research in Innovative Teaching & Learning*, *18*(2), 232–247.

Artigue, M. (2002). Learning mathematics in a CAS environment: The genesis of a reflection about instrumentation and the dialectics between technical and conceptual work. *International journal of computers for mathematical learning*, *7*(3), 245–274.

Bernardi, M. L., Capone, R., Faggiano, E., & Rocha, H. (2025). Generative AI in mathematics education: Pre-service teachers' knowledge and implications for their professional development. *International Journal of Mathematical Education in Science and Technology*, 1–18.

Bernardi, M. L., Capone, R., Faggiano, E., & Troilo, F. (2024). Teachers exploring the potential of generative AI in mathematics teaching. In E. Faggiano, A. Clark-Wilson, M. Tabach, & H.-G. Weigand (Eds.), *Proceedings of the 17th ERME Topic Conference MEDA4* (pp. 89–96). Università degli Studi di Bari Aldo Moro.

Capone, R., & Faggiano, E. (2024). Generative Artificial Intelligence scaffolding students' understanding of triple integrals. In *Proceedings of the 15th International Congress on Mathematical Education*, Sydney.

Chiu, T. K. (2024). Future research recommendations for transforming higher education with generative AI. *Computers and education: Artificial intelligence*, *6*, 100197.

Chiu, T. K., Moorhouse, B. L., Chai, C. S., & Ismailov, M. (2024). Teacher support and student motivation to learn with Artificial Intelligence (AI) based chatbot. *Interactive Learning Environments*, *32*(7), 3240–3256.

Hwang, G. J., Xie, H., Wah, B. W., & Gaševi , D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, *1*, 100001.

Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education*, 1–33.

Rabardel, P. (1995). *Les hommes et les technologies; approche cognitive des instruments contemporains.* Armand Colin

Troilo, F., Bernardi, M. L., Capone, R., & Faggiano, E. (2024). Generative AI as a tool to foster collective mathematical discussions in primary school classrooms. In E. Faggiano, A. Clark-Wilson, M. Tabach, & H.-G. Weigand (Eds.), *Proceedings of the 17th ERME Topic Conference MEDA4* (pp. 391–398). Università degli Studi di Bari Aldo Moro.

Trouche, L. (2004). Managing the complexity of human/machine interactions in computerized learning environments: Guiding students' command process through instrumental orchestrations. *International Journal of Computers for mathematical learning*, *9*(3), 281–307.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Williamson, B., Macgilchrist, F., & Potter, J. (2023). Re-examining AI, automation and datafication in education. *Learning, media and technology*, *48*(1), 1–5.

II.17
Formare i docenti per un uso consapevole dell'AI nella valutazione di una didattica ibrida

Claudio Palazzi
Dottorando in Teoria e Ricerca Educativa e Sociale,
Dipartimento di Scienze della Formazione, Università degli Studi Roma Tre, claudio.palazzi@uniroma3.it

Abstract

La presente ricerca analizza l'efficacia delle tecnologie immersive nell'educazione museale attraverso lo sviluppo e la validazione empirica del sistema MuSEd VR, una piattaforma di realtà virtuale che ricrea digitalmente gli spazi del Museo della Scuola e dell'Educazione dell'Università Roma Tre. Il sistema integra *avatar* di figure storiche della pedagogia italiana, animati da intelligenza artificiale secondo un'architettura *Retrieval-Augmented Generation* che garantisce coerenza storica e personalizzazione dell'esperienza educativa. Attraverso uno studio quasi-sperimentale con 2.524 partecipanti, la ricerca confronta l'esperienza immersiva con la visita museale tradizionale, documentando un miglioramento significativo nell'apprendimento oggettivo e nel coinvolgimento emotivo. I risultati attestano inoltre un rilevante effetto di democratizzazione culturale, con la riduzione delle barriere legate all'età e al capitale culturale pregresso. Lo studio propone il modello teorico dell'Apprendimento Immersivo Mediato, un *framework* che integra immersione tecnologica, mediazione pedagogica ed esperienza culturale per la progettazione di ambienti educativi ibridi.

Parole chiave: realtà virtuale educativa; Apprendimento Immersivo Mediato; intelligenza artificiale generativa; mediazione culturale; didattica museale.

This research analyzes the effectiveness of immersive technologies in museum education through the development and empirical validation of the MuSEd VR system, a virtual reality platform that digitally recreates the spaces of the Museum of School and Education at Roma Tre University. The system integrates avatars of historical figures in Italian pedagogy, animated by artificial intelligence using a Retrieval-Augmented Generation architecture that ensures historical coherence and personalization of the educational experience. Through a quasi-experimental study with 2,524 participants, the research compares the immersive experience with traditional museum visits, documenting significant improvement in objective learning and emotional engagement. Results also attest to a relevant cultural democratization effect, reducing barriers related to age and prior cultural capital.

The study proposes the Mediated Immersive Learning theoretical model, a framework integrating technological immersion, pedagogical mediation, and cultural experience for designing hybrid educational environments.

Keywords: educational virtual reality; Mediated Immersive Learning; generative artificial intelligence; cultural mediation; museum education.

1. Introduzione

La trasformazione digitale delle istituzioni educative e culturali rappresenta uno dei fenomeni più pervasivi e significativi del nostro secolo (Naqvi, 2017), processo che ha subito una drastica e inaspettata accelerazione a causa della pandemia di COVID-19. Questa contingenza storica ha reso improcrastinabile una riflessione approfondita sulle implicazioni pedagogiche, cognitive e sociali della digitalizzazione (Laurillard, 2012), portando alla luce le lacune esistenti nella letteratura scientifica, soprattutto per quanto riguarda la valutazione empirica e comparativa tra le esperienze didattiche tradizionali e quelle mediate dalle nuove tecnologie (Poce et al., 2020).

Nel contesto specifico della mediazione culturale museale, l'avvento delle tecnologie immersive solleva interrogativi fondamentali sulla loro capacità di arricchire l'esperienza educativa senza compromettere l'autenticità dell'incontro con il patrimonio culturale (Abdulwahed & Nagy, 2009). La ricerca del MuSEd VR ha evidenziato come la mera adozione tecnologica non garantisce automaticamente un miglioramento dei processi di apprendimento, rendendo necessaria un'analisi sistematica delle condizioni pedagogiche e progettuali che determinano l'efficacia degli ambienti virtuali educativi.

In risposta a questa esigenza di comprensione empirica e teorica, il presente contributo propone una valutazione sistematica dell'efficacia delle tecnologie immersive nell'educazione museale attraverso lo sviluppo e l'implementazione del sistema MuSEd VR. L'obiettivo della ricerca è duplice: da un lato, presentare il modello teorico dell'Apprendimento Immersivo Mediato, che articola le dimensioni fondamentali di un'esperienza educativa immersiva efficace. Dall'altro, documentare le evidenze empiriche emerse dalla validazione del sistema, con particolare attenzione al ruolo specifico dell'intelligenza artificiale generativa come componente della mediazione pedagogica all'interno dell'ambiente virtuale.

2. Quadro teorico: il modello dell'Apprendimento Immersivo Mediato

La trasformazione digitale della didattica e della mediazione culturale solleva interrogativi fondamentali che trascendono la mera adozione tecnologica per investire le dimensioni epistemologiche, pedagogiche ed etiche dell'esperienza educativa contemporanea. Nel percorso di ricerca è stato elaborato il modello teorico dell'Apprendimento Immersivo Mediato, un *framework* concettuale originale che articola tre dimensioni fondamentali interagenti sinergicamente.

2.1 Immersione tecnologica

La prima dimensione, denominata immersione tecnologica, riguarda il grado di coinvolgimento sensoriale e cognitivo generato dall'ambiente virtuale. Questa componente si fonda sulla teoria della presenza elaborata da Lombard e Ditton (Lombard & Ditton, 1997), che descrive quella peculiare sensazione di "essere là", di trovarsi realmente in un altro luogo e tempo, che costituisce l'essenza fenomenologica dell'esperienza immersiva. L'immersione tecnologica si realizza attraverso la riduzione delle distrazioni esterne e la creazione di un ambiente percettivo coerente che coinvolge i sensi del visitatore, facilitando processi attentivi focalizzati e un coinvolgimento emotivo intensificato rispetto alle modalità fruitive tradizionali (Cantatore, 2017).

2.2 Mediazione pedagogica

La seconda dimensione, definita mediazione pedagogica, implementa uno *scaffolding* digitale adattivo attraverso agenti di intelligenza artificiale che personalizzano dinamicamente il supporto secondo il livello di competenza dell'apprendente. Questa componente si ispira ai principi vygotskijani della zona di sviluppo prossimale e al concetto di *scaffolding* educativo (Fani & Ghaemi, 2011), riconoscendo che l'apprendimento efficace richiede un accompagnamento calibrato che fornisca supporto quando necessario e si ritiri progressivamente quando l'autonomia del discente lo consente. Nel contesto di MuSEd VR, la mediazione pedagogica è realizzata da *avatar* di figure storiche della pedagogia italiana animati da intelligenza artificiale generativa secondo un'architettura *Retrieval-Augmented Generation*. Questi *avatar* operano come mediatori culturali intelligenti capaci di sostenere dialoghi complessi con i visitatori, personalizzare le risposte in tempo reale, adattare i contenuti allo stile cognitivo in-

dividuale e modulare il livello di approfondimento in funzione delle conoscenze pregresse manifestate dall'interlocutore.

L'architettura *RAG* garantisce che le risposte degli *avatar* mantengano coerenza storica con il pensiero originale delle figure rappresentate. Quando, ad esempio, un visitatore chiede all'*avatar* che impersona Maria Montessori cosa pensasse delle tecnologie digitali, il sistema genera una risposta fedele ai principi pedagogici montessoriani, affermando che se uno strumento digitale permette al fruitore di esplorare autonomamente, di correggere i propri errori e di procedere secondo il proprio ritmo, allora potrebbe avere valore educativo ma che tale strumento non deve mai sostituire l'esperienza concreta con la realtà. Questa risposta esemplifica la capacità del sistema di mantenere fedeltà al pensiero storico mentre dialoga con interrogativi contemporanei che la figura rappresentata non ha potuto affrontare direttamente.

2.3 Esperienza culturale

La terza dimensione, l'esperienza culturale, preserva e attualizza consapevolmente i contenuti storici e pedagogici, evitando la visione tecnocratica dell'educazione. Questa componente sottolinea un principio fondamentale del modello: la tecnologia deve rimanere trasparente, al servizio dei contenuti culturali e dei valori pedagogici che si intendono trasmettere. La tentazione ricorrente in molti contesti di innovazione educativa è quella di cadere in un approccio tecnocratico che si concentra esclusivamente sulla tecnologia in sé, dimenticando il "perché" pedagogico che dovrebbe orientare ogni scelta progettuale. Nel sistema MuSEd VR, l'esperienza culturale è garantita dalla cura scientifica dei contenuti presentati negli ambienti virtuali e dalla configurazione degli *avatar* basata su fonti storiche certe. La tecnologia raggiunge il suo obiettivo quando diventa invisibile, quando il visitatore esce dall'esperienza riflettendo sulla profondità e l'attualità del pensiero pedagogico incontrato piuttosto che ricordando gli effetti speciali della virtualità.

L'efficacia del modello AIM risiede nell'integrazione sinergica delle tre componenti, poiché ciascuna dimensione potenzia le altre in un sistema dinamico e interdipendente. L'immersione tecnologica senza mediazione pedagogica genererebbe disorientamento e apprendimento superficiale, riducendo l'esperienza a una mera esplorazione passiva di ambienti virtuali. La mediazione pedagogica senza immersione tecnologica riprodurrebbe semplicemente le dinamiche della lezione tradizionale, perdendo le *affordance* uniche dell'ambiente immersivo quali la possibilità di osservare dettagli in modo ravvicinato, di muoversi nello

spazio tridimensionale e di esplorare al proprio ritmo. L'esperienza culturale senza mediazione efficace rimarrebbe astratta, incapace di tradursi in apprendimento significativo e trasferibile.

3. Il sistema MuSEd VR

MuSEd VR è una piattaforma di realtà virtuale che ricrea digitalmente gli spazi del Museo della Scuola e dell'Educazione dell'Università Roma Tre. L'esperienza interattiva si articola in quattro ambienti tematici dedicati a momenti cruciali della storia della pedagogia italiana, guidati da rispettivi *avatar* rappresentanti personaggi storici della pedagogia italiana come Duilio Cambellotti, Giuseppe Lombardo Radice, Maria Montessori e un noto personaggio della letteratura: Pinocchio. Tale rappresentazione consente ai visitatori di esplorare ricostruzioni fedeli di contesti educativi storici e di interagire con rappresentazioni virtuali di figure che hanno segnato l'evoluzione del pensiero pedagogico nazionale.

Dal punto di vista tecnologico, il sistema è fruibile attraverso visori *Meta Quest 3* e presenta una ricostruzione tridimensionale fedele degli spazi museali originali che consente la navigazione libera all'interno degli ambienti virtuali. I visitatori possono osservare i reperti con un livello di dettaglio impossibile nella fruizione tradizionale, avvicinandosi a documenti, oggetti e arredi scolastici storici per esaminarne particolari altrimenti inaccessibili. L'interazione vocale naturale con gli *avatar* rappresenta la componente più innovativa del sistema, permettendo libere conversazioni con le figure storiche rappresentate.

Gli *avatar* che guidano l'esperienza sono animati da un sistema di intelligenza artificiale generativa basato su architettura *Retrieval-Augmented Generation*. Questa architettura combina un modello linguistico di grandi dimensioni per la generazione di risposte naturali con un sistema di recupero documentale che attinge a fonti storiche primarie, il tutto vincolato da parametri di coerenza che assicurano fedeltà al pensiero originale delle figure rappresentate. Il risultato è un agente conversazionale capace di rispondere a domande aperte mantenendo accuratezza storica, riconoscendo i limiti della propria conoscenza e stimolando la riflessione critica. A differenza di un sistema linguistico generalista, gli *avatar* MuSEd VR sono configurati per rispondere coerentemente con i valori pedagogici della figura rappresentata, dichiarando esplicitamente quando una domanda esula dalle conoscenze documentate storicamente disponibili.

3.1 La valutazione di una didattica basata sull'AIM

La ricerca identifica nella formazione iniziale e continua degli insegnanti il fulcro strategico per realizzare il potenziale educativo nel valutare queste tecnologie. Tale preparazione può essere progettata secondo i principi del modello dell'Apprendimento Immersivo Mediato.

I futuri insegnanti possono sperimentare direttamente il modello *AIM*, praticando competenze didattiche in ambienti virtuali che simulano contesti scolastici realistici. Questa "pratica simulata immersiva" consiste in un'esperienza diretta di orchestrazione di ambienti di apprendimento complessi, mediati da tecnologie sofisticate. Ciò implica la necessità di formare gli insegnanti a un uso strategico e differenziato di questi strumenti, in primis l'*AI*, riconoscendo per quali obiettivi didattici offrono un reale valore aggiunto.

3.2 Valutazione di una didattica ibrida

La ricerca dal lato immersivo, con l'utilizzo della *Virtual Reality,* suggerisce un paradigma didattico intrinsecamente ibrido. Il modello *AIM* fornisce la cornice per la valutazione di una didattica che integri sinergicamente tecnologie immersive con metodologie tradizionali. Il futuro della fruizione didattica risiede nell'integrazione sinergica di fisico e digitale, valorizzando le specificità di ciascun mediatore culturale per creare percorsi educativi più ricchi, accessibili e democratici.

4. Metodologia

La validazione empirica del modello AIM è stata condotta attraverso uno studio che ha seguito un approccio *mixed-methods* concorrente, integrando simultaneamente dati quantitativi e qualitativi per ottenere una comprensione articolata dell'efficacia del sistema.

Il disegno quasi-sperimentale a gruppi indipendenti, con stratificazione del campione per controllare variabili confondenti, ha coinvolto complessivamente 2.524 partecipanti. Di questi, 1.443 hanno fruito dell'esperienza immersiva mediante visori *Meta Quest 3* della durata di 15 minuti, mentre 1.081 hanno effettuato la visita tradizionale al museo fisico, della medesima durata, costituendo il gruppo di controllo. Il campione è stato stratificato per età, *background* culturale e familiarità tecnologica, consentendo analisi differenziate dell'efficacia dell'intervento su sottogruppi con caratteristiche diverse.

Gli strumenti di misurazione hanno incluso test di apprendimento oggettivo strutturati su tre livelli cognitivi secondo la tassonomia di Bloom, permettendo di valutare sia la memorizzazione di informazioni sia l'elaborazione concettuale a livelli superiori. Il coinvolgimento emotivo è stato rilevato attraverso scale *Likert* con punteggi da 1 a 10, integrate da analisi qualitativa delle domande aperte per brevi testimonianze dei partecipanti. L'accettazione tecnologica è stata misurata attraverso il *System Usability Scale* secondo il *Technology Acceptance Model* (Money & Turner, 2004), consentendo di valutare la percezione di usabilità e accessibilità del sistema indipendentemente dall'esperienza pregressa con tecnologie immersive.

4.3 Risultati

I risultati della ricerca documentano differenze significative tra le due modalità di fruizione, attestando una superiorità dell'esperienza immersiva mediata in molteplici indicatori. Per quanto riguarda l'apprendimento oggettivo, nella modalità VR si è registrata una media dell'84,4% di risposte corrette ai test di verifica, rispetto al 35,6% ottenuto dal gruppo che ha effettuato la visita al museo tradizionale. Questa differenza di 48,8 punti percentuali rappresenta un *effect size di Cohen's d* pari a 2,79, un valore che può essere considerato molto elevato e che indica un impatto sostanziale dell'intervento. È particolarmente significativo che nell'esperienza VR non si sia osservata differenziazione sostanziale tra i tre livelli cognitivi della tassonomia di Bloom, suggerendo che l'immersione adeguatamente mediata facilita l'elaborazione a livello di memorizzazione ed anche a livelli concettuali superiori che richiedono comprensione, applicazione e analisi.

I risultati relativi al coinvolgimento emotivo confermano la superiorità dell'esperienza immersiva. L'89,7% dei partecipanti all'esperienza VR ha riportato un coinvolgimento emotivo intenso, con punteggi medi di 8,5 su una scala da 1 a 10. Nel gruppo che ha effettuato la visita tradizionale, questa percentuale scende al 52,3%, con punteggi medi di 5,2 su 10. Le analisi qualitative delle testimonianze hanno evidenziato autentici momenti di risonanza emotiva con le figure storiche rappresentate dagli *avatar*, con i partecipanti che riferiscono una percezione di autenticità nel dialogo e una connessione personale con il patrimonio culturale che trascende la mera acquisizione di informazioni.

Un risultato di particolare rilevanza sociale riguarda l'effetto livellante della tecnologia VR sulle differenze individuali, configurando il sistema come potente strumento di democratizzazione culturale. I partecipanti con età superiore ai

cinquanta anni nel gruppo VR hanno ottenuto *performance* del 68,4%, risultati comparabili a quelli dei partecipanti giovani che hanno raggiunto il 71,2%. Questo dato sfida il persistente mito del *digital divide* generazionale, dimostrando che quando l'esperienza è pedagogicamente ben progettata la tecnologia diventa accessibile indipendentemente dall'età. Analogamente, la correlazione tra livello di istruzione pregresso e apprendimento, marcata nel museo tradizionale, si è ridotta significativamente nell'esperienza VR, indicando che il sistema è in grado di compensare differenze nel capitale culturale di partenza.

L'analisi dell'usabilità attraverso il *System Usability Scale* ha confrontato utenti esperti e novizi di realtà virtuale. Gli esperti hanno ottenuto un punteggio *SUS* medio di 84,1, mentre i novizi si sono attestati a 81,7, con un gap di soli 2,4 punti. Il t-test per campioni indipendenti ha restituito un valore t pari a 1,97 con p superiore a 0,05, indicando che la differenza non raggiunge la significatività statistica. L'esperienza pregressa in VR non influenza dunque in modo significativo la percezione di usabilità, confermando che l'interfaccia MuSEd VR rispetta i principi di usabilità universale e consente una effettiva democratizzazione dell'accesso alla tecnologia immersiva.

5. Discussione

I risultati empirici validano il modello dell'Apprendimento Immersivo Mediato, confermando che l'integrazione sinergica delle tre dimensioni produce effetti educativi significativamente superiori rispetto all'esperienza tradizionale. I dati suggeriscono che la superiorità dell'esperienza immersiva risiede nella qualità della progettazione pedagogica che la orienta e nella mediazione che accompagna l'utente nel suo percorso di scoperta.

L'architettura *RAG* degli *avatar* garantisce che l'intelligenza artificiale operi effettivamente come *scaffolding* adattivo, calibrando il supporto sulle esigenze cognitive ed emotive del singolo visitatore. L'assenza di differenziazione tra livelli cognitivi nei risultati VR indica che l'immersione, quando accompagnata da mediazione pedagogica efficace, facilita la memorizzazione e l'elaborazione concettuale. Questo dato supporta l'interpretazione dell'intelligenza artificiale implementata nel sistema come mediatore culturale intelligente capace di attivare processi cognitivi complessi attraverso il dialogo personalizzato.

La ricerca non suggerisce tuttavia la sostituzione dell'esperienza fisica con quella virtuale. L'esperienza del museo tradizionale conserva un valore unico legato all'autenticità dell'oggetto materiale, alla fisicità della visita e alle dinamiche sociali che si sviluppano nella fruizione condivisa degli spazi culturali. I risultati

delineano piuttosto le fondamenta per un paradigma didattico intrinsecamente ibrido, nel quale tecnologie immersive e metodologie tradizionali si integrano sinergicamente valorizzando le specificità di ciascun mediatore culturale. Il modello AIM fornisce un *framework* operativo per progettare questa integrazione, riconoscendo che il futuro della fruizione didattica museale risiede nella combinazione strategica di fisico e digitale per creare percorsi educativi più ricchi, accessibili e democratici.

Lo studio presenta alcune limitazioni che devono essere considerate nell'interpretazione dei risultati. Il campione, pur ampio, è stato reclutato in un contesto geografico specifico concentrato nell'area di Roma, limitando la universalità dei risultati ad altri contesti territoriali e culturali. L'effetto novità della tecnologia VR potrebbe aver influenzato positivamente le risposte emotive, sebbene i dati sull'apprendimento oggettivo suggeriscano che gli effetti documentati trascendano il mero entusiasmo iniziale. Il *follow-up* a lungo termine della ritenzione delle conoscenze non è stato ancora adottato, rendendo necessarie ulteriori indagini per valutare la persistenza degli effetti nel tempo.

6. Conclusioni

La presente ricerca documenta empiricamente l'efficacia del sistema MuSEd VR e del modello teorico dell'Apprendimento Immersivo Mediato. I dati attestano che le tecnologie immersive, quando progettate secondo i principi del *framework* AIM che integra immersione tecnologica, mediazione pedagogica ed esperienza culturale, producono miglioramenti significativi nell'apprendimento oggettivo e nel coinvolgimento emotivo dei visitatori.

Particolarmente rilevante è l'effetto di democratizzazione culturale documentato dalla ricerca: la tecnologia VR, quando adeguatamente progettata, riduce le barriere legate all'età e al capitale culturale pregresso, rendendo l'esperienza educativa accessibile al pubblico tradizionalmente escluso o svantaggiato nella fruizione museale tradizionale. L'intelligenza artificiale generativa, implementata attraverso l'architettura *RAG* negli *avatar* conversazionali, si configura come componente essenziale della mediazione pedagogica, operando come *scaffolding* adattivo che personalizza l'esperienza senza compromettere la coerenza storica e culturale dei contenuti. L'intelligenza artificiale nel contesto di MuSEd VR amplifica l'esperienza culturale autentica attraverso un dialogo personalizzato che rispetta la fedeltà storica e stimola la riflessione critica.

Le tecnologie digitali, quando guidate da una solida visione pedagogica e orientate da un impegno etico verso l'accessibilità e la qualità culturale, possono

amplificare e arricchire l'esperienza educativa tradizionale, realizzando il potenziale di un apprendimento autenticamente immersivo e culturalmente consapevole.

Riferimenti bibliografici

Abdulwahed, M., & Nagy, Z. (2009). *Applying Kolb's experiential learning cycle for laboratory education*.

Cantatore, L. (2017). Primo: Leggere / a cura di Lorenzo Cantatore: Per un'educazione alla lettura. In *Primo: Leggere: Per un'educazione alla lettura*. Conoscenza.

Fani, T., & Ghaemi, F. (2011). Implications of Vygotsky's Zone of Proximal Development (ZPD) in Teacher Education: ZPTD and Self-scaffolding. *Procedia - Social and Behavioral Sciences*, *29*, 1549-1554. https://doi.org/10.1016/j.sbspro.2011.11.396

Laurillard, D. (2012). *Teaching as a Design Science: Building Pedagogical Patterns for Learning and Technology* (1ª ed.). Routledge.

Lombard, M., & Ditton, T. (1997). At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication*, *3*(2), 0-0. https://doi.org/10.1111/j.1083-6101.1997.tb00072.x

Money, W., & Turner, A. (2004). Application of the technology acceptance model to a knowledge management system. *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, 9. https://doi.org/10.1109/HICSS.2004.1265573

Naqvi, S. (2017). *Towards an Integrated Framework: Use of Constructivist and Experiential Learning Approaches in ICT-Supported Language Learning*. https://consensus.app/papers/towards-an-integrated-framework-use-of-constructivist-and-naqvi/9e744060a40f5c11a4254e779e664bed/

Poce, A., Amenduni, F., Re, M. R., Medio, C. D., & Norgini, A. (2020). Correlations among natural language processing indicators and critical thinking sub-dimensions in HiEd students. *Form@re - Open Journal per La Formazione in Rete*, *20*(3), Articolo 3. https://doi.org/10.13128/form-9902

II.18
Valutazione dell'impatto dell'Intelligenza Artificiale nella formazione docente: una Scoping Review per un modello pedagogico-organizzativo

Alessandro Barca
Professore Associato in PAED-02/B - Link Campus University - a.barca@unilink.it[1]

Abstract

Il contributo mappa criticamente la letteratura sull'impatto dell'intelligenza artificiale nei percorsi di formazione iniziale e in servizio degli insegnanti, mediante una scoping review guidata da PRISMA-ScR. La sintesi tematica, letta attraverso il framework AgID di valutazione d'impatto, evidenzia uno sbilanciamento verso indicatori percettivi e tecnico pedagogici, a fronte di scarse evidenze su trasferimento in pratica, ricadute organizzative e governance etica. Si delinea un modello pedagogico organizzativo ecologico-sistemico, fondato su indicatori misti e longitudinali.

Parole chiave: intelligenza artificiale; formazione docente; scoping review; valutazione d'impatto; modello ecologico-sistemico.

This paper maps the evidence on the impact of artificial intelligence in pre service and in service teacher education through a PRISMA ScR guided scoping review. Using the AgID impact assessment framework as an interpretive grid, the analysis highlights a strong emphasis on self reported, techno pedagogical outcomes, alongside limited evidence on classroom transfer, organisational change and ethical governance. An ecological pedagogical organisational model is proposed, grounded in mixed and longitudinal indicators.

Keywords: artificial intelligence; teacher education; scoping review; impact assessment; ecological-systems model.

1 Alessandro Barca è Professore Associato (SSD PAED-02/B; S.C. 11/D2) presso Link Campus University. È Direttore di una collana editoriale di fascia A, Edizioni TAB, referee di riviste scientifiche e collane editoriali e fa parte di numerosi comitati scientifici nazionali ed internazionali; è stato, inoltre, membro di progetti nazionali (FRC e PRA). Ha partecipato come relatore e chair a numerosi convegni nazionali e internazionali ed è autore di pubblicazioni scientifiche nazionali e internazionali.

1. Introduzione: contesto e rilevanza

La società globale sta attraversando una transizione digitale di ordine epocale, caratterizzata non solo dalla diffusione di nuove tecnologie, ma da una riorganizzazione sistemica delle relazioni sociali, dei modelli economici e degli stessi processi cognitivi. Questo paradigma, spesso definito "onlife" (Floridi, 2015) per la fusione tra esperienza online e offline, impone una ridefinizione delle competenze necessarie per l'agire consapevole e la piena cittadinanza. Il sistema educativo, in tale scenario, non è un mero recettore passivo di innovazione, ma un attore centrale chiamato a formare individui dotati di *digital literacy*, pensiero critico e capacità di adattamento a contesti complessi e in continua evoluzione (OECD, 2019). Si configura, di fatto, un processo sistemico di innovazione tecnologica, accelerato dal Piano Nazionale per l'Innovazione 2025 e dal Decennio Digitale UE, con l'adozione diffusa di Intelligenza Artificiale (AI), robotica e piattaforme cloud nel quadro della PA Digitale 2026. Nonostante progressi significativi, come il miglioramento di 30 punti nelle competenze digitali degli alunni all'ICILS 2023 rispetto al 2018, persistono divari strutturali: solo il 45,8% della popolazione tra i 16 e i 74 anni possiede competenze digitali di base, con l'Italia al 22° posto UE, e solo il 62% delle scuole ha connessioni adeguate, mentre il 25% delle scuole pubbliche integra strumenti IA. Tali disparità, accentuate dal 66% degli insegnanti privo di formazione adeguata su IA e dal 46% che teme svantaggi per gli studenti senza accesso tecnologico, richiedono un intervento educativo mirato per colmare il gap infrastrutturale e generazionale.

Proprio per questo, la figura del docente trascende il tradizionale ruolo di trasmettitore di conoscenze per configurarsi, in termini più pregnanti, quale moltiplicatore critico di competenze (Gavosto, 2022). Questo costrutto sottolinea una duplice funzione: da un lato, il docente è un moltiplicatore, il cui livello di padronanza e capacità di integrazione didattica delle tecnologie determina un effetto leva sull'apprendimento e sullo sviluppo di abilità degli studenti; dall'altro, è un filtro critico, o dovrebbe esserlo, essenziale per guidare gli alunni a interrogare, valutare e utilizzare in modo etico e produttivo gli strumenti digitali, contrastando un approccio meramente acritico o consumistico (Mishra & Koehler, 2006).

La formazione docente diviene, quindi, il nodo strategico decisivo per l'integrazione efficace e sostenibile dell'Intelligenza Artificiale (AI) nei processi educativi. L'AI, infatti, non costituisce una semplice estensione degli strumenti digitali già noti, ma introduce elementi di discontinuità radicale: algoritmi opachi, processi decisionali automatizzati, personalizzazione dell'apprendimento

su larga scala e nuove questioni etiche relative a *bias*, equità e responsabilità (Zawacki-Richter et al., 2019).

Investire in una formazione docente robusta, *evidence-based* e focalizzata sulle dimensioni tecnico-pedagogiche, organizzative ed etiche dell'AI non è dunque una scelta opzionale, ma una condizione necessaria per tradurre le promesse della trasformazione digitale in reali miglioramenti degli esiti educativi, per promuovere un'inclusione autentica e per costruire una scuola capace di preparare le future generazioni alle sfide di un mondo sempre più plasmato dagli algoritmi (European Commission, 2022).

Alla luce di tale contesto, il presente saggio si propone di sistematizzare il panorama conoscitivo esistente riguardante l'impatto dell'AI nella formazione iniziale e continua degli insegnanti. La letteratura in questo dominio, seppur in rapida crescita, appare frammentata e caratterizzata da approcci valutativi eterogenei.

Al fine di organizzare la riflessione e fornire una base strutturata per l'analisi, lo studio adotta come cornice teorica di riferimento il *framework* per la valutazione di impatto dell'IA sviluppato dall'Agenzia per l'Italia Digitale (AgID, 2025). Questo modello, concepito originariamente per il settore pubblico, offre una griglia multidimensionale particolarmente adatta a informare e ispirare un modello di valutazione specifico per il contesto educativo. La scelta di questo *framework* non è meramente strumentale ma concettuale: esso rappresenta uno dei primi tentativi istituzionali nazionali di superare una valutazione riduzionista, focalizzata esclusivamente su parametri di efficienza tecnica, a favore di un'analisi olistica degli effetti generati dalle tecnologie algoritmiche (AgID, 2025).

Il *framework* AgID propone una valutazione d'impatto articolata lungo tre dimensioni interdipendenti, qui riadattate come griglia ermeneutica per la formazione docente.

- La dimensione tecnico-pedagogica, che consente di analizzare l'acquisizione di competenze digitali specifiche (come la *prompt engineering* o l'analisi dei dati educativi) e, soprattutto, la loro effettiva integrazione e rielaborazione in pratiche didattiche innovative e coerenti con i modelli pedagogici di riferimento.
- La dimensione organizzativa, cruciale per indagare come l'introduzione dell'IA influisca sui processi scolastici, sulla ridefinizione dei ruoli professionali, sulla cultura collaborativa dell'istituto e sulla gestione delle risorse, spostando l'attenzione dall'individuo al sistema.
- La dimensione etica e valutativa, che impone di scrutinare le implicazioni in termini di equità, trasparenza (*explainability*), responsabilità (*accountability*)

e rispetto dell'autonomia umana, sia nell'uso didattico degli strumenti sia nella loro eventuale applicazione nei processi di valutazione degli apprendimenti.

La pertinenza di questo *framework* risiede nella sua capacità di collegare la valutazione micro (le competenze del docente) a quella meso (l'organizzazione scolastica) e macro (i principi etici e di *policy*). Esso fornisce così un linguaggio comune e una tassonomia operativa per mappare, classificare e interpretare le evidenze sparse nella letteratura, consentendo di superare la frammentazione degli studi esistenti. Il suo utilizzo in questo saggio non è quindi passivo, ma critico e adattivo: si propone di testare la trasferibilità di uno schema valutativo nato in ambito amministrativo alla realtà complessa della formazione con la specificità dei processi educativi e della professionalità docente, contribuendo al contempo ad arricchire lo stesso *framework* con le suggestioni provenienti dalla ricerca pedagogica.

2. Approccio metodologico e prime evidenze

Per rispondere alla natura esplorativa e sistematizzante degli obiettivi di ricerca, è stata adottata la metodologia della *Scoping Review* (Arksey, O'Malley, 2005). Al fine di garantire rigore, trasparenza e replicabilità del processo di revisione, la procedura è stata guidata dal protocollo PRISMA-ScR (*Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews*) proposto da Tricco e colleghi (2018). Questo garantisce trasparenza e riproducibilità attraverso checklist per reporting, enfatizzando inclusione di fonti eterogenee senza valutazione critica di qualità e svolgendo una funzione cartografica critica: identificare i costrutti principali, chiarire le definizioni operative in un dominio semanticamente fluido e, soprattutto, mappare le lacune (*gap*) conoscitive in modo sistematico, al fine di fornire una base solida per future indagini empiriche e lo sviluppo di *framework* valutativi (Fig. 1).

L'indagine documentale è stata condotta interrogando tre database bibliografici multidisciplinari e specialistici, selezionati per la loro rilevanza nel campo delle scienze dell'educazione e delle tecnologie digitali.

Il processo di screening, condotto in doppio cieco con due ricercatori indipendenti, ha applicato criteri di inclusione/esclusione predefiniti su titolo, abstract e testo completo, garantendo l'affidabilità della selezione del corpus documentale finale, come dettagliato nello schema sintetico (Tab. 1). Sono stati inclusi studi empirici, revisioni della letteratura, framework teorici e documenti

di *policy* pubblicati in inglese e italiano, senza limiti temporali rigidi ma con una focalizzazione preferenziale sul quinquennio 2020-2025 per cogliere le evoluzioni più recenti.

Fase	Azione Condotta	Strumento/Protocollo impiegato	Output principale
Identificazione	Definizione obiettivi e domanda di ricerca esplorativa	Analisi contesto e letteratura grigia	Protocollo di ricerca predefinito
Ricerca	Interrogazione sistematica di database bibliografici	Stringhe di ricerca booleane. *Database: Google Scholar*, ERIC, *ScienceDirect*	Set iniziale di record (n = 13.679)
Screening	Applicazione criteri di inclusione/esclusione	Doppia revisione cieca; diagramma di flusso PRISMA	Set finale di studi inclusi (n = 12)
Estrazione	Analisi contenutistica e codifica dei dati	Griglia di analisi basata sul Framework AgID (2025)	Matrice di dati codificati per dimensione
Sintesi	Mappatura concettuale e analisi tematica.	Sintesi narrativa e tabellare	Identificazione modelli, costrutti e *gap*

Tab. 1 – Percorso metodologico della Scoping Review condotta

Per garantire la massima pertinenza e copertura tematica, la strategia di ricerca è stata costruita mediante l'impiego di una stringa booleana complessa, basata sui tre concetti cardinali dello studio: l'intervento, la popolazione e *l'outcome*. L'utilizzo combinato dell'operatore *AND* tra i blocchi concettuali, e dell'operatore *OR* all'interno dei blocchi per sinonimi e traduzioni (italiano/inglese), ha massimizzato l'efficacia di recupero dei documenti. Le stringhe di ricerca che combinano termini relativi al dominio tecnologico sono state: *"artificial intelligence"*, *"machine learning"*, al target professionale *"teacher training"*, *"professional development"*, e all'oggetto di indagine *"impact assessment"*, *"evaluation framework"*.

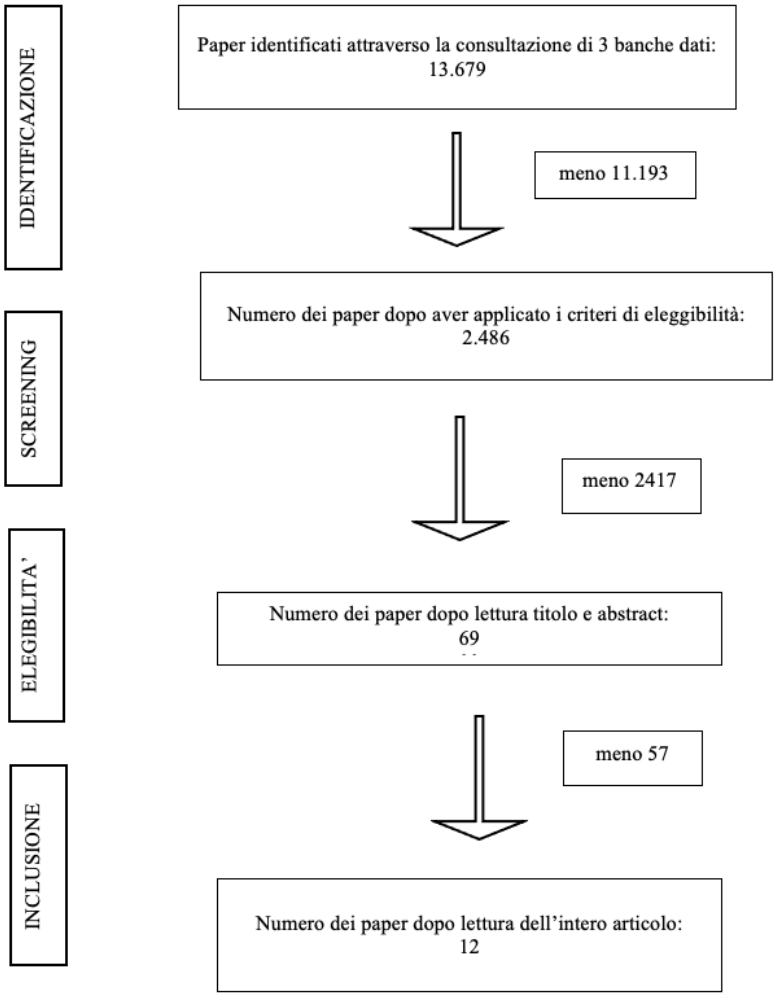


Fig. 1 – Diagramma di Flusso della revisione svolta

L'analisi tematica dei documenti selezionati è stata condotta utilizzando una griglia di codifica derivata dal *Framework* AgID (2025), adattandone le tre dimensioni: tecnico-pedagogica, organizzativa, etica e valutativa come categorie ermeneutiche per una sintesi strutturata (Tab. 2). Questo approccio ha consentito non solo di organizzare i risultati, ma di valutare criticamente il focus e l'ampiezza della letteratura esistente attraverso una lente multidimensionale consolidata.

Articolo	Anno	Tipo	Focus principale	Framework	Dimensione AgID	Gap rilevato
Teachers' trust in AI-powered educational technology (Nazaretsky et al.)	2022	Empirico	Fiducia e accettazione tecnologica dei docenti	Trust-based (TAM/TPB)	Tecnico-ped. + Etico-val.	Resistenza passiva vs uso critico/consapevole
Teacher training in the age of AI: impact on AI literacy (Lademann et al.)	2025	Pre-post (N=436)	AI Literacy, attitudini e fiducia nell'integrazione	Kirkpatrick Level 2 (Apprendimento)	Tecnico-ped.	Discrepanza tra alfabetizzazione teorica e fiducia pratica
AI in Teacher Education: Unlocking Potential (Chandra)	2025	Analisi Qualitativa	ITS, apprendimento adattivo e analisi educativa	NEP 2020 / UNESCO	Tutte (Sistemica)	Lacune etiche e assenza di politiche coordinate
Evaluation of a Program to Enhance online Lecturers' Digital Competence (Romero-García et al.)	2025	Pre-post (N=89)	Competenze digitali e GenAI in ambito universitario	TPACK + GenAI (INTEF)	Tecnico-ped.	Scarso trasferimento nella pratica (solo 23%)
Challenges and best practices in training teachers (Aljemely)	2024	Systematic Review	Formazione sistemica vs episodi isolati	Ecologico-Sistemico	Organizzativa	Formazione episodica e mancanza di visione strategica
Artificial Intelligence in Primary Education: Systematic Review (Bagdonaitė & Dagienė)	2025	PRISMA Review (94 studi)	Stato dell'arte e PD nella scuola primaria	AI-TPACK	Tutte (Sistemica)	Frammentazione dei modelli di valutazione e gap di evidenze
In search of AI literacy in teacher education (Sperling et al.)	2024	Scoping Review	Concettualizzazione della AI Literacy nella formazione iniziale	Scoping Review Framework	Tecnico-ped.	Frammentazione dei gap etici e pratici nella letteratura
AI in teaching and teacher professional development (Tan et al.)	2025	Systematic Review	Opportunità di sviluppo professionale e competenze necessarie	AI Competence Framework	Tutte (Sistemica)	Necessità di integrazione sistematica e non episodica
Generative AI in teacher education (Prilop et al.)	2025	Mixed Methods	Percezioni di potenziali trasformativo e natura triade AI Literacy	Triadic AI Literacy	Etico-valutativa	Assenza di procedure di audit algoritmico e bias di confidenza
A new framework for teachers' professional development (Sancar et al.)	2021	Conceptual / Review	Superamento della formazione tradizionale "a cascata"	Integrated PD Framework	Organizzativa	Disconnessione tra formazione teorica e cambiamento nelle pratiche d'aula
The Impact of AI Teaching on Teaching Quality (Wang)	2025	Empirico (SEM)	Effetto dell'IA mediato da motivazione e competenze docenti	AI-TPACK Theory	Tecnico-ped.	Effetto diretto IA sovrastimato; il ruolo della mediazione umana è centrale
Delphi study for AI literacy assessment (Laupichler et al.)	2023	Delphi Study	Validazione item per misurare la AI Literacy in non esperti	Socio-cognitive Literacy	Etico-valutativa	Mancanza di strumenti di misura standardizzati e validati

Tab. 2 – Mappatura dei paper identificati

L'applicazione di questa griglia analitica ha prodotto tre principali evidenze. In primo luogo, è stata possibile una mappatura critica dei modelli di valutazione prevalenti nella letteratura, i cui limiti e potenzialità sono sintetizzati nella Tab. 3. Emerge una tripartizione tra:

– modelli centrati sulla conoscenza docente, come il *Technological Pedagogical Content Knowledge* (TPACK) (Mishra, Koehler, 2006), efficaci nel concettualizzare l'integrazione ma spesso valutati tramite auto-rapporti;
– modelli di valutazione della formazione, come i *Quattro Livelli* di Kirkpatrick e Kirkpatrick (2006), ampiamente utilizzati ma raramente applicati oltre i livelli di reazione e apprendimento individuale;
– *framework* etici prescrittivi (es., UNESCO, 2024; European Commission, 2022), che forniscono principi ma scarseggiano di indicatori operativi misurabili.

Modello/Framework	Focus principale	Potenzialità	Limitazioni emerse
TPACK (Mishra & Koehler, 2006)	Conoscenza integrata docente (tecnologica, pedagogica, disciplinare)	Supera la visione strumentale; concettualizza l'integrazione complessa	Difficoltà di misurazione oggettiva; forte affidamento su *self-report*
Livelli di Kirkpatrick (2006)	Valutazione a quattro livelli degli esiti della formazione	Fornisce una struttura gerarchica e progressiva per la valutazione	Letteratura carente sui livelli 3 (comportamento) e 4 (risultati); focus individualistico
Framework Etici (come quelli riportati in UNESCO, 2024)	Principi guida per un uso equo e responsabile dell'IA	Definiscono un orizzonte valoriale imprescindibile per l'implementazione	Natura prescrittiva; assenza di indicatori misurabili e di protocolli di valutazione concreta

Tab. 3 – Mappatura e Analisi Critica dei Modelli di Valutazione Identificati

In secondo luogo, l'analisi ha identificato i costrutti e gli indicatori privilegiati dalla letteratura. Prevale un netto squilibrio verso la dimensione tecnico-pedagogica individuale, con costrutti ricorrenti quali l'autoefficacia tecnologica percepita e l'intenzione comportamentale d'uso, misurati quasi esclusivamente tramite strumenti percettivi (questionari). Gli indicatori osservativi del trasferimento in pratica e, soprattutto, quelli relativi all'impatto organizzativo (come ad esempio il cambiamento nei processi decisionali collegiali) ed etico-valutativo (*audit* degli algoritmi) risultano estremamente rari e frammentari.

Questo sbilanciamento conduce alla terza e più significativa evidenza: l'esistenza di un marcato *gap evidence-based*. La letteratura mostra una dissociazione sistematica tra la misurazione dell'acquisizione di abilità tecnologiche e la documentazione della loro traduzione in modelli pedagogici sostenibili e miglioramenti organizzativi misurabili (Zawacki-Richter et al., 2019). Manca una ricerca longitudinale che tracci l'evoluzione delle pratiche didattiche e, soprattutto, sono assenti strumenti valutativi capaci di cogliere l'impatto ecologico-sistemico dell'AI, ovvero le interconnessioni tra il cambiamento individuale del docente, la trasformazione della cultura scolastica e l'adeguamento delle *policy*. Questo vuoto evidenzia l'inadeguatezza degli approcci valutativi esistenti e giustifica la necessità di un nuovo modello pedagogico-organizzativo, informato da una prospettiva ecologica che colleghi organicamente tutte e tre le dimensioni del *framework* AgID.

L'analisi tematica condotta sul *corpus* definitivo di 12 studi selezionati attraverso il protocollo PRISMA-ScR conferma in modo empirico la frammentazione del campo di ricerca e ne articola le criticità fondamentali, fornendo una base solida per un salto paradigmatico nella valutazione d'impatto. Una prima consistente categoria di studi si concentra quasi esclusivamente sulla dimensione tecnico-pedagogica individuale, applicando prevalentemente il modello TPACK o teorie dell'accettazione tecnologica (TAM). Questi lavori, spesso basati su *survey* e auto-rapporti, forniscono dati quantitativi su autoefficacia, intenzioni d'uso e percezione delle competenze. Tuttavia questo approccio soffre di un *deficit* di validità ecologica: misura il costrutto psicologico ma non ne verifica il trasferimento in azioni didattiche osservabili e sostenibili nel tempo. Studi come quello di Käser e Alexandron (2024), che tenta di integrare micro-analisi di interazione in ambienti di simulazione, rappresentano un'eccezione promettente ma isolata. Una seconda categoria, più ridotta, comprende studi di caso organizzativo, come il lavoro etnografico di Schmidt et al. (2021), su un distretto scolastico canadese e l'analisi di processo di Alerasoul et al. (2022) in Giappone. Questi contributi, qualitativi e longitudinali, iniziano a esplorare la dimensione organizzativa, evidenziando come l'adozione dell'IA alteri dinamiche di potere, richieda nuove figure di supporto e generi tensioni nella cultura professionale. Tuttavia, il loro limite, riconosciuto dagli stessi autori, è la difficoltà di generalizzazione e la mancanza di un quadro sistematico per confrontare contesti diversi. Infine, una serie di *framework* concettuali e linee guida, dall'UNESCO (2024) a contributi accademici come quello di Borenstein e Howard (2023), affrontano la dimensione etica. Pur essendo fondamentali, questi documenti rimangono largamente prescrittivi. Come notato nell'analisi comparativa, esiste un divario sostanziale, un *implementation gap,* tra i principi enunciati (equità, trasparenza) e la loro traduzione in indicatori valutativi operativi applicati nei programmi di formazione docente. Non un singolo *paper* presenta un *audit* algoritmico condotto dagli stessi docenti a seguito di una formazione. La mappatura di questo *corpus* non rivela solo squilibri tematici, anche un problema epistemologico più profondo. La predominanza di disegni di ricerca trasversali (*cross-sectional*), di strumenti di misurazione percettiva e di contesti decontestualizzati (corsi brevi e isolati) indica un'adesione a un paradigma di ricerca di stampo positivista e riduzionista. Questo paradigma è intrinsecamente inadeguato a catturare la natura complessa, adattiva e situata del fenomeno in esame: la trasformazione della professionalità docente in risposta a tecnologie algoritmiche che, a loro volta, evolvono e interagiscono con un ecosistema organizzativo dinamico.

La quasi totale assenza nel *corpus* di studi longitudinali *mixed-methods* che colleghino dati psicometrici, osservazioni etnografiche in aula e analisi di docu-

menti organizzativi (PTOF, verbali) è sintomatica. Misurare separatamente le variabili (competenza, uso, impatto) per poi cercare correlazioni lineari fallisce nel cogliere le proprietà emergenti e le retroazioni (*feedback loops*) tipiche dei sistemi complessi. Ad esempio, nessuno studio analizza come un cambiamento nelle pratiche valutative di un docente, abilitato dall'AI, possa innescare una revisione del regolamento di dipartimento, che a sua volta rafforza o inibisce nuove pratiche. È questa diagnosi metodologica, emersa direttamente dall'analisi critica dei paper *estratti*, a costituire la giustificazione scientifica più solida per la proposta di un modello ecologico-sistemico. Tale modello non è una semplice aggiunta di dimensioni alla griglia AgID, ma propone un cambio di unità di analisi: dall'individuo-docente o dal singolo strumento tecnologico, al sistema insegnante-tecnologia-contesto organizzativo nella sua interazione dinamica.

3. Discussione e proposta di modello pedagogico-organizzativo

3.1 Un approccio ecologico-sistemico per interpretare i risultati

I risultati emersi dalla *scoping review* non possono essere interpretati in modo isolato, ma richiedono una cornice teorica in grado di coglierne la complessità interdipendente. Un approccio ecologico-sistemico (Bronfenbrenner, 1979), riadattato al contesto organizzativo-scolastico, fornisce la lente ideale per leggere criticamente le evidenze. Questo approccio concepisce l'innovazione tecnologica non come un evento puntuale i diversi livelli del sistema educativo e li trasforma: dal microsistema della singola aula e della relazione docente-studente-strumento, al mesosistema delle interazioni dipartimentali e della cultura scolastica, fino all'esosistema delle politiche e delle norme istituzionali (Senge et al., 2012). Alla luce di questa prospettiva, la mappatura dei modelli di valutazione evidenzia una criticità strutturale: la predominanza di strumenti ancorati a una logica individuale e lineare, incapaci di cogliere la natura relazionale e contestuale dell'impatto. Modelli come i primi due livelli di Kirkpatrick (2006), reazione e apprendimento, sebbene utili, rischiano di cristallizzare una valutazione che si arresta alla sfera psicologica e cognitiva del singolo formatore, trascurando il successivo passaggio cruciale. È pertanto imperativo sviluppare strumenti valutativi che orientino espressamente l'analisi verso il trasferimento nelle pratiche professionali, come indicato dal terzo livello di Kirkpatrick, relativo al comportamento, e, soprattutto, verso l'impatto sull'organizzazione scolastica nel suo complesso, ponendo attenzione ai risultati, secondo il livello 4. Ciò implica misurare non solo se un docente sappia usare uno strumento di AI, ma come questa

competenza modifichi la progettazione didattica collaborativa, le dinamiche di mentoring tra colleghi, l'allocazione delle risorse tecnologiche della scuola e, in definitiva, la capacità organizzativa di apprendere e innovarsi in modo sistematico (Fullan, 2007).

3.2 Sintesi critica della letteratura

L'analisi sistematica del corpus selezionato (Tab. 2) evidenzia un panorama scientifico in rapida evoluzione, un quadro ormai maturo ma ancora strutturalmente incompleto della formazione dei docenti sull'Intelligenza Artificiale, nel quale emergono convergenze forti su criticità sistemiche, pedagogiche ed etico-valutative.

Il quadro complesso che emerge può essere definito come il "paradosso della competenza": a fronte di incrementi misurabili (+24%) nella AI Literacy teorica (Lademann et al., 2025), si riscontra un trasferimento nelle pratiche d'aula estremamente limitato, stimato solo al 23% (Romero-García et al., 2025). Tale evidenza suggerisce la necessità di far evolvere il tradizionale framework TPACK verso un modello "AI-TPACK" più robusto (Bagdonaité & Dagiené, 2025), capace di trasformare la confidenza percepita in un uso critico e consapevole (Nazaretsky et al., 2022). In questo scenario, la mediazione umana si conferma il fattore determinante; analisi empiriche basate su "Modelli di Equazioni Strutturali" (SEM) dimostrano infatti che l'AI non possiede un'efficacia intrinseca sulla qualità della didattica, ma agisce solo se mediata dall'expertise del docente e dalla motivazione degli studenti (Wang, 2025), riposizionando l'insegnante come filtro pedagogico essenziale contro ogni logica di automazione tecnologica.

Parallelamente, i risultati spostano l'attenzione dalla dimensione individuale a quella organizzativa, denunciando l'obsolescenza dei modelli formativi "episodici" o "a cascata" a favore di una visione strategica e sistemica (Sancar et al., 2021; Aljemely, 2024) che identifichi nella scuola un'organizzazione capace di apprendere costantemente (Tan et al., 2025). Permane tuttavia un preoccupante "vuoto operativo" sul piano etico e metodologico: la ricerca rileva infatti l'assenza di procedure di audit algoritmico nelle pratiche docenti (Prilop et al., 2025), una carenza di strumenti di misurazione standardizzati per l'AI Literacy (Laupichler et al., 2023) e una frammentazione nelle policy di indirizzo (Chandra, 2025; Sperling et al., 2024). Complessivamente, le evidenze suggeriscono che i limiti attuali dell'integrazione dell'AI non siano di natura tecnologica, bensì epistemica e organizzativa: senza una sintesi sistemica tra alfabetizzazione, sviluppo professionale continuo e governance istituzionale, l'uso dell'AI rischia di rima-

nere un'innovazione prevalentemente retorica piuttosto che realmente trasformativa. Di fatto, risulta necessario colmare il gap relativo alla valutazione dell'impatto a lungo termine (livelli 3 e 4 di Kirkpatrick), integrare l'audit etico nei percorsi di sviluppo professionale e promuovere modelli di leadership distributiva capaci di sostenere l'innovazione come reale cambiamento culturale.

3.3 La proposta centrale: un modello a competenze trasformative ed evidence-based

Partendo dall'analisi ecologico-sistemica e dai *gap* evidenziati, si formalizza la proposta di un modello pedagogico-organizzativo per la valutazione d'impatto dell'IA nella formazione dei docenti. Questo modello ha come finalità primaria quella di fornire una guida *evidence-based* a *policy-maker*, dirigenti scolastici e formatori, spostando il baricentro dalla misurazione di competenze tecniche specifiche alla valutazione di competenze trasformative (UNESCO, 2024).

Il modello integra e rielabora le tre dimensioni del *framework* AgID (2025) in una struttura dinamica e circolare, illustrata nella Fig. 2, dove:

- la dimensione tecnico-pedagogica diviene il motore delle competenze trasformative individuali, ossia la capacità del docente di ri-progettare la propria azione didattica, di valutare criticamente gli output algoritmici e di agire come designer di esperienze di apprendimento abilitate dall'AI;
- la dimensione organizzativa è il terreno di coltura necessario, che valuta la capacità della scuola di sostenere questa trasformazione attraverso leadership distribuita, tempi strutturati per la collaborazione e una visione strategica condivisa dell'innovazione;
- la dimensione etica e valutativa funge da impalcatura di governo, garantendo che il processo sia guidato da principi di equità, trasparenza e responsabilità, integrati sia nelle pratiche didattiche sia nei protocolli istituzionali.

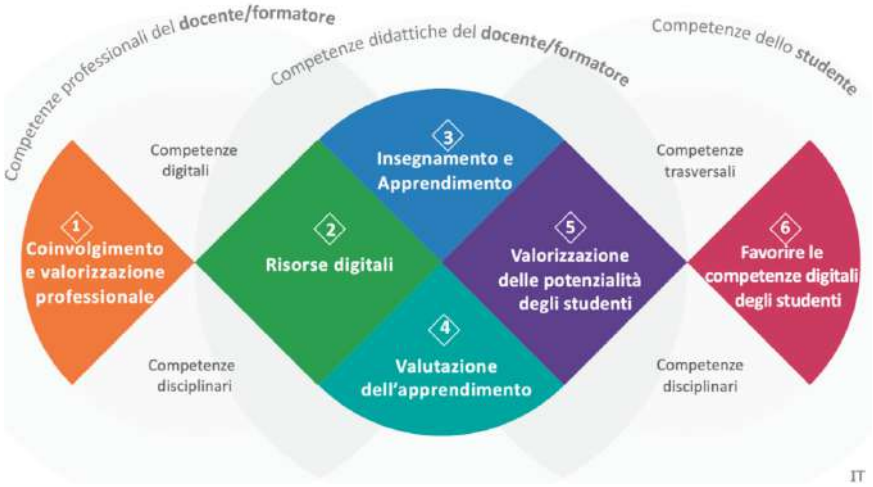


Fig. 2 – Modello pedagogico-organizzativo integrato a competenze trasformative

Il modello propone quindi una griglia operativa di indicatori misti (qualitativi e quantitativi) per ciascuna dimensione, progettati per superare la dipendenza esclusiva dai *self-report*. Per la dimensione tecnico-trasformativa, oltre ai questionari, si propongono l'analisi di portfolio digitali dei docenti e l'osservazione sistematica di lezioni per cogliere l'integrazione pedagogica dell'AI. Per la dimensione organizzativa si considerano indicatori come la frequenza e la qualità delle riunioni di co-progettazione dipartimentale, l'evoluzione dei PTOF in relazione al digitale e i sondaggi sul clima di innovazione. Per la dimensione etica, infine, si propongono l'adozione formale di linee guida per l'uso dell'AI nella scuola e la documentazione di casi di discussione critica su *bias* algoritmici nei consigli di classe. Questa proposta non intende essere prescrittiva, ma rappresenta un *framework* flessibile e adattabile, che pone al centro la trasformazione sistemica e sostenibile come vero indicatore di successo della formazione docente nell'era dell'AI, colmando il vuoto *evidence-based* rilevato nella letteratura.

4. Conclusioni

Il presente contributo ha perseguito l'obiettivo di sistematizzare la letteratura esistente sull'impatto dell'Intelligenza Artificiale nella formazione docente, adottando come griglia ermeneutica il *framework* per la valutazione di impatto sviluppato da AgID (2025). La sua adozione critica si è rivelata fondamentale per

superare una visione riduzionista e tecnocentrica, consentendo di organizzare l'analisi lungo tre dimensioni interdipendenti (tecnico-pedagogica, organizzativa, etico-valutativa) e di evidenziarne le connessioni sistemiche (AgID, 2025; Zawacki-Richter et al., 2019). Questo approccio multidimensionale ha permesso di mappare non solo i modelli valutativi prevalenti, ma soprattutto di diagnosticare con precisione i limiti epistemologici che li caratterizzano. L'analisi condotta attraverso la *Scoping Review* conferma in modo inequivocabile un marcato *gap evidence-based* nel dominio di ricerca esaminato. La letteratura appare dominata da studi trasversali e strumenti di *self-report*, focalizzati su costrutti individuali come l'autoefficacia percepita, mentre risultano estremamente carenti ricerche longitudinali e strumenti in grado di documentare il trasferimento delle competenze in pratiche didattiche sostenibili, l'impatto organizzativo e l'integrazione operativa dei principi etici (Borenstein & Howard, 2021).

4.1 Implicazioni pratiche e linee future

Le evidenze emerse hanno implicazioni dirette per i *policy-maker* e le istituzioni formative. È imperativo abbandonare modelli formativi episodici e incentrati esclusivamente sulle *skill* tecniche, a favore di percorsi strutturati e longitudinali che affrontino organicamente le tre dimensioni del *framework*. I decisori politici sono chiamati a promuovere e finanziare programmi che integrino la formazione tecnico-pedagogica con un forte focus sullo sviluppo organizzativo (attraverso la definizione di figure di sistema, come gli AI *coordinator*) e sulla sperimentazione di protocolli etici operativi, quali *l'audit* algoritmico partecipato (Schmidt et al., 2021; European Commission, 2025). Inoltre, è necessario favorire una cultura collaborativa e fornire tempi e risorse dedicati alla co-progettazione e alla riflessione critica sulle pratiche, affinché l'innovazione divenga patrimonio del sistema-scuola e non del singolo individuo (Fullan, 2007; Senge et al., 2012).

Per colmare il vuoto conoscitivo e metodologico rilevato, la ricerca futura deve compiere un deciso salto paradigmatico. Si rendono necessari:

– studi longitudinali di tipo *mixed-methods* che combinino dati psicometrici, osservazioni etnografiche in aula e analisi documentale per tracciare la co-evoluzione delle competenze docenti, delle pratiche didattiche e delle strutture organizzative;
– lo sviluppo e la validazione di strumenti valutativi multidimensionali ed *evidence-based*, ispirati al modello ecologico-sistemico qui abbozzato, che superino la dipendenza dagli auto-rapporti attraverso l'uso di portfolio digitali,

protocolli osservativi e indicatori organizzativi (es., analisi dei PTOF, verbali di dipartimento).
– ricerche-azione partecipate finalizzate a tradurre i *framework* etici prescrittivi (es., UNESCO) in strumenti concreti e auditivi, testandone l'applicabilità e l'efficacia nei contesti scolastici reali.

Solo abbracciando questa logica della complessità e investendo in un approccio valutativo rigoroso e olistico sarà possibile trasformare la promessa dell'AI in un reale moltiplicatore di qualità educativa, equità e innovazione sostenibile per l'intero sistema scolastico.

Riferimenti bibliografici

AgID - Agenzia per l'Italia Digitale. (2025). *Appendice Linee Guida IA nella PA - Valutazione impatto*. https://www.agid.gov.it/sites/agid/files/2025-02/Linee_Guida_-adozione_IA_nella_PA.pdf

Alerasoul, S.A., Afeltra, G., Hakala, H., Minelli, E., & Strozzi, F. (2022). Organisational learning, learning organisation, and learning orientation: An integrative review and framework. *Human Resource Management Review*, Volume 32, Issue 3, September 2022, 100854. https://doi.org/10.1016/j.hrmr.2021.100854

Aljemely, Y. (2024). Challenges and best practices in training teachers to utilize artificial intelligence: a systematic review. *Frontiers in Education*, 9, 1470853. https://doi.org/10.3389/feduc.2024.1470853

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19-32. https://doi.org/10.1080/1364557032000119616

Bagdonaité, J., & Dagiené, V. (2025). Artificial Intelligence in Primary Education: A Systematic Literature Review 2020–2025. *Informatics in Education*, 24(4), 697-736. https://doi.org/10.15388/infedu.2025.24

Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI and ethics*, *1*(1), 61-65. https://doi.org/10.1007/s43681-020-00002-7

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge (MA): Harvard University Press. https://doi.org/-10.2307/j.ctv26071r6

Chandra, A., (2025). Artificial Intelligence in Teacher Education: Unlocking Potential, Navigating Challenges and Shaping the Future. *The Academic International Journal of Multidisciplinary Research. Volume 3 | Issue 4 | April 2025 ISSN: 2583-973X.*

Commissione Europea. (2025). Relazione di monitoraggio del settore dell'istruzione e della formazione 2025. relazione di monitoraggio del settore dellistruzione-NC0125134ITN (1).pdf

European Commission. (2022). *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators*. Publications Office of the European Union. ethical guidelines on the use of artificial intelligence-NC0722649ENN.pdf

Floridi, L. (2009). *The Onlife Manifesto. Being Human in a Hyperconnected Era.* Milano: Springer Open.

Fullan, M. (2007). *The new meaning of educational change* (4th ed.). New York: Teachers College Press.

Gavosto, A. (2022). *La scuola bloccata.* Roma-Bari: Laterza.

Käser, T., & Alexandron, G. (2024). Simulated Learners in Educational Technology: A Systematic Literature Review and a Turing-like Test. *Int J Artif Intell Educ* 34 (2), 545–585. https://doi.org/10.1007/s40593-023-00337-2

Kirkpatrick, D.L., & Kirkpatrick, J.D. (2006). *The four levels of training evaluation* (3rd ed.). Oakland: Berrett-Koehler Publishers.

Lademann, J., Henze, J., Honke, N., Wollny, C., & Becker-Genschow, S. (2025). Teacher training in the age of AI: impact on AI literacy and teachers' attitudes. *Computers and Society (cs.CY); Artificial Intelligence (cs.AI). 2507.03011v1*

Laupichler, M.C., Aster, A., & Raupach, T. (2023). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. *Computers and Education: Artificial Intelligence 4, 100126.* https://doi.org/-10.1016/j.caeai.2023.100126

Mishra, P., & Koehler, M.J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017-1054. DOI: 10.1111/j.1467-9620.2006.00684.x

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI powered educational technology and a professional development program to improve it. *British Journal of Educational Technology,* 53, 914-931. https://doi.org/-10.1111/bjet.13232Teachers_trust_in_AI-powered_educational_technolo.pdf

OECD. (2019). *Going Digital: Shaping Policies, Improving Lives*. OECD Publishing. https://doi.org/10.1787/9789264312012-en

Prilop, C.N., Mah, D.-K., Jacobsen, L.J., Weber, K.E., & Hoya, F. (2025). Generative AI in teacher education: Educators' perceptions of transformative potentials and the triadic nature of AI literacy explored through AI-enhanced methods. *Computers and Education: Artificial Intelligence, 9, 100471.* https://doi.org/10.1016/-j.caeai.2025.100471

Romero-García, C., Buzón-García, O., & Parra González, E. (2025). Evaluation of a Program to Enhance online Lecturers' Digital Competence: TPACK and Artificial Intelligence, Online Learning, 29(4), 109-132. https://doi.org/10.-24059/olj.v29i4.5016

Sancar, R., Atal, D., & Deryakulu, D. (2021). A new framework for teachers' professional development. *Teaching and Teacher Education, Volume 101, May 2021, 103305*. https://doi.org/10.1016/j.tate.2021.103305

Schmidt, L., Finnerty Mutlu, A.N., Elmore, R., Olorisade, B.K., Thomas, J., & Higgins, J.P.T. (2021). Data extraction methods for systematic review (semi)automation: Update of a living systematic review. *F1000Research*, *10*, 401. https://doi.org/10.12688/f1000research.51117.3

Senge, P., Cambron-McCabe, N., Lucas, T., Smith, B., Dutton, J., & Kleiner, A. (2012). *Schools that learn: A fifth discipline fieldbook for educators, parents, and everyone who cares about education.* New York City: Crown Business.

Sperling, K., Stenberg, C.-J., McGrath, C., Åkerfeldt, A., Heintz, F., & Stenliden, L. (2024). In search of artificial intelligence (AI) literacy in teacher education: A scoping review. *Computers and Education Open* 6 (2024) 100169. https://doi.org/-10.1016/j.caeo.2024.100169

Tan, X., Cheng, G., & Ling, M.H. (2025). Artificial intelligence in teaching and teacher professional development: A systematic review. *Computers and Education Open* 8, 100355. https://doi.org/10.1016/j.caeai.2024.100355

Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D.J., Horsley, T., Weeks, L., Hempel, S., Akl, E.A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M.G., Garritty, C., Lewin, S., & Straus, S.E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of internal medicine*, *169*(7), 467-473. https://doi.org/10.7326/M18-0850

UNESCO. (2024). *AI and education: Guidance for policy-makers*. https://unesdoc.unesco.org/ark:/48223/pf0000376709/PDF/376709eng.pdf.multi ai_and_education_-_guidance_for_policy-makers.pdf

Wang, J. (2025). The Impact of AI Teaching on Teaching Quality: The Mediating Effect of Student Motivation and Teacher Expertise. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)20(1).* https:// orcid .org/ 0009 -0008 -1197 -7112

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 39, 1–27. https://doi.org/10.1186/s41239-019-0171-0

II.19
Docenti in formazione davanti all'IA: percezioni, pratiche e questioni etiche della valutazione

Silvia Artusi
Dottoranda INFO-01/A
Università degli Studi Internazionali di Roma, silvia.artusi@unint.eu
Rossana Pia Laccone
Dottoranda PSIC-02/A, Università degli Studi Internazionali di Roma
rossana.laccone@unint.eu
Marco Romano
Professore Associato di Informatica INFO-01/A
Università degli Studi Internazionali di Roma
marco.romano@unint.eu

Abstract

L'ingresso dell'Intelligenza Artificiale generativa nei contesti educativi sta ridefinendo in modo significativo le pratiche valutative, sollevando questioni didattiche, etiche e professionali di particolare rilevanza, soprattutto nella formazione iniziale degli insegnanti. Il contributo presenta un'esperienza formativa realizzata all'interno di un modulo universitario di Didattica Digitale, finalizzata a esplorare le percezioni, le pratiche e le criticità che emergono quando docenti in formazione sono chiamati a confrontare il proprio giudizio valutativo con quello prodotto da sistemi di Intelligenza Artificiale generativa. Attraverso la valutazione di compiti autentici, inizialmente condotta dai docenti e successivamente affidata a un modello generativo, l'esperienza ha reso visibili convergenze e discrepanze tra valutazione umana e automatizzata, attivando processi di riflessività professionale. I risultati evidenziano come l'IA sia percepita come un utile strumento di supporto per gli aspetti formali del feedback, ma inadeguata a cogliere la complessità pedagogica, relazionale e contestuale dell'apprendimento. Il contributo sottolinea la necessità di integrare percorsi di AI literacy critica nella formazione degli insegnanti e di promuovere modelli valutativi ibridi, nei quali il giudizio professionale umano resti centrale e insostituibile.

Parole chiave: Intelligenza Artificiale; valutazione; agire riflessivo; formazione degli insegnanti; AI literacy.

The introduction of generative Artificial Intelligence into educational contexts is significantly reshaping assessment practices, raising pedagogical, ethical, and professional issues of particular relevance, especially in initial teacher education. This paper presents a training experience carried out within a university module on Digital Didactics, aimed at exploring the perceptions, practices, and critical issues that emerge when pre-service teachers are asked to compare their own evaluative judgment with that produced by generative Artificial Intelligence systems. Through the assessment of authentic tasks—initially conducted by the teachers and subsequently entrusted to a generative model—the experience made visible convergences and discrepancies between human and automated evaluation, fostering processes of professional reflexivity. The findings highlight that AI is perceived as a useful support tool for the formal aspects of feedback, yet inadequate in capturing the pedagogical, relational, and contextual complexity of learning. The study emphasizes the need to integrate critical AI literacy pathways into teacher education and to promote hybrid assessment models in which human professional judgment remains central and irreplaceable.

Keywords: Artificial Intelligence; assessment; reflective practice; teacher training; AI literacy.

1. Introduzione

L'arrivo dell'Intelligenza Artificiale generativa nel mondo dell'istruzione sta producendo trasformazioni significative non solo nelle modalità di insegnamento, ma anche nella concezione della valutazione. La crescente disponibilità di modelli capaci di generare testi complessi, analisi linguistiche e feedback dettagliati in tempi brevi viene spesso interpretata come una risposta efficace al carico di tempo e di impegno mentale che caratterizza il lavoro valutativo degli insegnanti. In questa prospettiva, l'AI appare come soluzione tecnica a un problema strutturale, promettendo efficienza e standardizzazione (González-Calatayud et al., 2021; Bulut et al., 2024). Le ricerche pedagogiche più recenti evidenziano tuttavia alcune ambiguità. Il rischio è che la valutazione venga ridotta a un'operazione prevalentemente computazionale, focalizzata sull'output e meno attenta alla complessità dei processi di apprendimento e alla dimensione relazionale dell'azione educativa (Ranieri, 2024; Fiorucci & Bevilacqua, 2024). Accelerare la valutazione può significare rendere più veloce un processo che, per sua natura, richiede tempo, interpretazione e riflessione. La lentezza, spesso percepita come inefficienza, costituisce invece uno dei suoi elementi formativi. Per i docenti in formazione iniziale questa tensione assume un valore particolare. In tale fase, apprendere a valutare coincide con la costruzione dell'identità professionale. Il

confronto con l'IA non rappresenta solo un'acquisizione tecnica, ma un'occasione per interrogare criticamente il significato del giudizio educativo. L'IA può così funzionare come "specchio epistemologico", rendendo visibili criteri e assunzioni implicite che orientano l'agire valutativo: che cosa significa valutare? Quale spazio resta all'interpretazione e alla responsabilità quando il giudizio è mediato da un algoritmo (Pillera, 2023; Perla & Vinci, 2024)? La letteratura mostra come i sistemi generativi risultino particolarmente efficaci nell'analisi di aspetti formali e linguistici, soprattutto quando tali dimensioni sono esplicitamente richieste nei prompt, ma meno affidabili nel cogliere la profondità concettuale, la creatività e il percorso di apprendimento sotteso al prodotto finale (Bewersdorff et al., 2023; Cohn et al., 2024). È in questo scarto che si colloca la dimensione formativa della valutazione, intesa come pratica interpretativa e relazionale. Una delega integrale all'IA rischierebbe di ridurre il lavoro interpretativo che consente al docente di comprendere l'allievo come soggetto situato (Brembilla, 2025). Il presente contributo analizza un'esperienza formativa realizzata in un modulo universitario di Didattica Digitale, finalizzata a esplorare percezioni, pratiche e questioni etiche emergenti dal confronto tra giudizio umano e output di un sistema di Intelligenza Artificiale generativa, in un setting valutativo intenzionalmente progettato. Attraverso la valutazione di compiti autentici, inizialmente condotta dai docenti e successivamente affidata a un modello generativo tramite prompt elaborati dagli stessi insegnanti, l'esperienza ha reso visibili convergenze e discrepanze difficilmente rilevabili in un uso non problematizzato della tecnologia. In questo contesto, l'IA è stata utilizzata non come sostituto, ma come dispositivo critico capace di attivare processi di riflessività professionale (Creati & Cuccaro, 2025; Fiorucci & Bevilacqua, 2025).

2. IA e processi valutativi: quadro concettuale e criticità

La letteratura più recente sottolinea come la diffusione dell'Intelligenza Artificiale nei contesti educativi renda necessario sviluppare forme di alfabetizzazione digitale non solo tecniche, ma critiche, capaci di integrare dimensioni cognitive, operative ed etico-sociali (Ranieri, 2024). Le tecnologie educative non sono strumenti neutrali: incorporano logiche di funzionamento e modelli impliciti di apprendimento che orientano, talvolta in modo poco visibile, pratiche formative e valutative. In questa prospettiva si collocano le riflessioni della media education, che invitano a interrogare criticamente algoritmi e sistemi automatizzati, promuovendo un uso consapevole e trasparente delle tecnologie (Panciroli et al., 2020). L'impiego dell'Intelligenza Artificiale nei processi valutativi

rappresenta uno degli ambiti più controversi di tale trasformazione. Studi recenti evidenziano come i sistemi generativi producano feedback articolati e formalmente accurati, soprattutto quando i criteri sono esplicitati nel prompt, risultando spesso percepiti come più sistematici rispetto a quelli dei pari (Usher, 2025). Ciò ha alimentato l'idea dell'IA come valido supporto alla valutazione, in termini di rapidità e standardizzazione. Tuttavia, emergono criticità ricorrenti: tendenza alla sovrastima delle prestazioni e focalizzazione sugli aspetti più osservabili del prodotto, soprattutto in assenza di informazioni contestuali e longitudinali. In tali condizioni, la valutazione automatizzata appare meno adeguata nell'interpretare il senso educativo dell'apprendimento e le traiettorie di sviluppo degli studenti. Accanto ai limiti tecnico-pedagogici, la letteratura evidenzia preoccupazioni etiche legate all'opacità algoritmica, alla gestione dei dati e al rischio di delega di decisioni educative a sistemi automatici (Zawacki-Richter et al., 2019). In questo quadro, la centralità del docente risulta decisiva: la valutazione resta un atto interpretativo che richiede sensibilità pedagogica e conoscenza del contesto. Come osservano Perla e Vinci (2024), un uso formativo dell'IA è possibile solo all'interno di processi valutativi ibridi, nei quali la tecnologia supporta e amplifica il giudizio umano senza sostituirlo.

3. Metodologia

L'esperienza è stata realizzata nell'ambito di un corso universitario rivolto a docenti in formazione iniziale, all'interno di un modulo di didattica digitale. La collocazione in questo contesto risponde all'esigenza di osservare come l'Intelligenza Artificiale venga interpretata e utilizzata da insegnanti che, pur avendo esperienza scolastica, si trovano in una fase di rielaborazione critica del proprio ruolo e delle proprie pratiche valutative. La formazione iniziale costituisce infatti uno spazio privilegiato per attivare processi di riflessività professionale, nei quali le tecnologie emergenti possono essere interrogate anche sul piano pedagogico ed etico (Ranieri, 2024). Il gruppo era composto da quindici docenti di differenti ambiti disciplinari e ordini di scuola secondaria, eterogenei per anni di servizio e abitudini valutative. Tale varietà è stata considerata una risorsa metodologica, poiché ha consentito di osservare come uso e percezione dell'IA siano influenzati da esperienza, contesto disciplinare e concezioni implicite di valutazione. Il gruppo ha quindi funzionato come micro-osservatorio delle pratiche e rappresentazioni presenti nella scuola. L'impianto metodologico è stato esplorativo e riflessivo, con l'obiettivo di mettere in dialogo giudizio umano e giudizio automatizzato, senza assumere una prospettiva gerarchica. Per processo valutativo

ibrido si intende un dispositivo in cui la valutazione resta in capo al docente, mentre l'IA genera un feedback parallelo successivamente discusso e rielaborato.

L'attività si è articolata in più fasi. In una prima fase è stata proposta un'introduzione teorica sui principi di funzionamento dei modelli di Intelligenza Artificiale generativa, con attenzione al ruolo del prompt come mediazione tra intenzionalità umana e output algoritmico. Questa fase ha evidenziato la non neutralità della tecnologia e la dipendenza delle risposte dalle istruzioni fornite. Sono stati inoltre discussi i limiti della valutazione automatizzata rispetto alla capacità di cogliere aspetti contestuali, intenzionali e relazionali (González-Calatayud et al., 2021; Zawacki-Richter et al., 2019). Nella seconda fase, ciascun docente ha selezionato un compito autentico svolto da uno studente della propria classe, valutandolo autonomamente con giudizio motivato e breve feedback scritto secondo i criteri abituali. Questa fase ha reso espliciti criteri e priorità valutative spesso impliciti. Successivamente, i compiti sono stati sottoposti a un modello di Intelligenza Artificiale generativa tramite prompt elaborati dai docenti, richiedendo feedback e suggerimenti coerenti con criteri esplicitati (chiarezza, coerenza, correttezza linguistica, aderenza alla consegna). Non sono state fornite informazioni sistematiche sul percorso dello studente né richieste valutazioni longitudinali: il confronto si è concentrato sul prodotto e sui criteri dichiarati. La struttura del prompt prevedeva: (i) consegna; (ii) testo dello studente; (iii) criteri del docente; (iv) richiesta di feedback argomentato. Un esempio di prompt è il seguente: «Valuta il seguente elaborato sulla base dei criteri indicati (chiarezza espositiva, coerenza argomentativa, correttezza linguistica, aderenza alla consegna). Fornisci un feedback argomentato e suggerimenti operativi di miglioramento, evitando di assegnare un voto numerico». Il confronto tra feedback umano e automatico ha costituito il nucleo riflessivo dell'esperienza, rendendo visibili convergenze e discrepanze (Usher, 2025). Eventuali focalizzazioni dell'output sono state interpretate anche alla luce delle istruzioni contenute nei prompt, distinguendo tra limiti del modello e scelte incorporate nella richiesta. La fase conclusiva ha previsto una riflessione metacognitiva e la raccolta dei dati tramite questionario misto (domande chiuse e aperte) volto a rilevare percezioni su affidabilità, potenzialità e rischi dell'IA. Poiché alcuni limiti teorici erano già stati discussi, il questionario ha mirato a rilevarne la rielaborazione alla luce dell'esperienza concreta (Perla & Vinci, 2024). Non sono stati raccolti dati pre-intervento: l'obiettivo non era misurare variazioni di atteggiamento, ma analizzare come le rappresentazioni teoriche venissero rinegoziate attraverso il confronto con l'output del modello.

4. Risultati

Dall'analisi incrociata dei questionari e dei materiali prodotti emergono alcune tendenze ricorrenti trasversali ai diversi profili professionali coinvolti. Una prima evidenza riguarda il riconoscimento diffuso delle capacità dell'Intelligenza Artificiale nel fornire un'analisi puntuale degli aspetti formali dei testi. I docenti sottolineano come, nel setting adottato e in relazione ai criteri esplicitati nei prompt, l'IA risulti efficace nell'individuare errori grammaticali, imprecisioni sintattiche, incoerenze testuali e registri inadeguati. Il feedback generato viene spesso descritto come linguisticamente raffinato e strutturato, talvolta più "elegante" rispetto a quello prodotto nella quotidianità scolastica, soprattutto in condizioni di carico lavorativo elevato. Tale percezione è coerente con la letteratura, che evidenzia come i sistemi generativi tendano a fornire commenti estesi e formalmente accurati quando la richiesta è focalizzata sul prodotto (Usher, 2025).

Accanto a questo riconoscimento emerge una valutazione critica sul piano pedagogico. Nelle condizioni operative dell'esperienza – assenza di informazioni sul percorso pregresso e prompt centrati sul prodotto – il modello mostra difficoltà, nel setting descritto e in assenza di informazioni longitudinali, nel cogliere elementi centrali dell'agire valutativo, quali intenzionalità dello studente, percorso di apprendimento, grado di autonomia e profondità concettuale. L'IA appare dunque competente nel "leggere" il testo, ma, nelle condizioni adottate, meno adeguata nel restituirne il significato educativo complessivo. A titolo esemplificativo, in un compito di argomentazione disciplinare l'IA ha valorizzato correttezza sintattica e completezza espositiva, attribuendo un giudizio elevato, mentre il docente ha evidenziato superficialità concettuale e limitata problematizzazione. La discrepanza ha riguardato non la forma, ma la profondità interpretativa.

In diversi casi è stata inoltre rilevata una tendenza alla sovrastima delle prestazioni, con giudizi complessivamente più alti rispetto a quelli dei docenti. Tale tendenza viene ricondotta anche alla focalizzazione del prompt su correttezza e completezza, in linea con studi che mostrano come l'IA assegni valutazioni più generose quando opera su indicatori formali (Usher, 2025).

Un ulteriore elemento riguarda la percezione di distanza tra feedback automatico e umano. Pur giudicato corretto e ben formulato, il commento dell'IA è descritto come impersonale e poco situato, nel setting descritto e in assenza di informazioni contestuali. Alcuni docenti osservano che il feedback "suona bene", ma non restituisce il contesto in cui il compito è stato prodotto; altri affermano che l'IA "vede la correttezza, ma non lo sforzo", mostrando, nelle condizioni

operative adottate, una limitata capacità di valorizzare progressi e fragilità. Viene inoltre segnalata l'incapacità del modello, nel setting adottato, di riconoscere il valore formativo di tentativi parziali o soluzioni non convenzionali.

Sul piano etico-professionale emergono preoccupazioni legate al rischio di riduzione della dimensione relazionale della valutazione e di delega impropria della responsabilità docente. Alcuni partecipanti evidenziano il pericolo di una standardizzazione impersonale, incapace, nelle condizioni operative adottate, di cogliere la complessità dell'allievo e del contesto. Tali riflessioni si collocano in continuità con le criticità segnalate dalla letteratura in merito all'opacità algoritmica e alla perdita di controllo umano nei processi decisionali (Zawacki-Richter et al., 2019; Mele & Gentile, 2023).

5. Discussione

I risultati dell'esperienza confermano che l'Intelligenza Artificiale può rappresentare un supporto ai processi valutativi nelle condizioni operative adottate, ma solo se inserita all'interno di un quadro intenzionalmente pedagogico e criticamente orientato. Da un lato, i docenti riconoscono le potenzialità operative dell'IA, in particolare in termini di rapidità, sistematicità e ricchezza del feedback, soprattutto con riferimento agli aspetti esplicitamente richiesti nei prompt, che possono facilitare una revisione più accurata dei testi e sostenere una maggiore esplicitazione dei criteri valutativi. In questo senso, l'IA appare come uno strumento in grado di alleggerire alcune componenti del lavoro valutativo, in particolare quelle più ripetitive e formali, contribuendo a una maggiore efficienza del processo.

Dall'altro lato, l'esperienza mette in evidenza come la valutazione non possa essere ridotta a un'operazione di analisi testuale o di attribuzione di punteggi. La valutazione educativa si configura come un atto complesso, situato e relazionale, che richiede comprensione del contesto, conoscenza dello studente e capacità di interpretare il significato dell'apprendimento nel suo sviluppo temporale. Tali dimensioni, nel setting adottato e in assenza di informazioni longitudinali sul percorso dello studente, non sono risultate pienamente accessibili ai sistemi di Intelligenza Artificiale utilizzati, anche in ragione della focalizzazione dei prompt prevalentemente sul prodotto.

La distanza riscontrata tra la valutazione umana e quella automatica evidenzia la necessità di sviluppare una competenza professionale specifica nell'uso dell'IA. Non si tratta semplicemente di saper utilizzare uno strumento, ma di acquisire una postura riflessiva che consenta di comprenderne limiti, implicazioni e con-

dizioni di utilizzo. In questo senso, la formazione iniziale degli insegnanti emerge come uno spazio privilegiato per elaborare un rapporto critico con l'IA, evitando sia atteggiamenti di rifiuto aprioristico sia forme di delega acritica. Come sottolineato da diversi studi, l'IA può sostenere il lavoro docente solo se integrata all'interno di processi valutativi ibridi, nei quali il giudizio umano rimane centrale e insostituibile (Perla & Vinci, 2024).

Nel setting adottato, il processo valutativo ibrido si è articolato in quattro fasi:

(1) formulazione autonoma del giudizio professionale del docente;
(2) esplicitazione dei criteri valutativi nel prompt;
(3) generazione di un feedback parallelo da parte dell'IA;
(4) confronto critico e rielaborazione finale del giudizio da parte del docente.

L'ibridazione non consiste dunque in una delega del processo, ma in una copresenza di giudizi che mantiene la responsabilità valutativa in capo all'insegnante.

Un aspetto particolarmente rilevante emerso dall'esperienza riguarda il valore formativo del confronto tra valutazione umana e valutazione automatica. La discrepanza tra i due giudizi ha costretto i docenti a esplicitare criteri, priorità e aspettative che spesso restano impliciti nella pratica quotidiana. In questo senso, la dimensione riflessiva non è attribuibile all'IA in quanto tale, ma al confronto critico tra il giudizio professionale del docente e l'output algoritmico, nelle condizioni operative adottate. Piuttosto che sostituire il docente, l'IA ha agito come uno "specchio critico", capace di rendere discutibili le scelte valutative e di stimolare una maggiore consapevolezza rispetto alle responsabilità educative connesse alla valutazione.

6. Conclusioni

Nel complesso, i risultati indicano che i docenti in formazione, nel contesto dell'esperienza descritta, adottano un atteggiamento prudente e riflessivo nei confronti dell'Intelligenza Artificiale. Ne riconoscono le potenzialità operative, in particolare la rapidità e la qualità formale del feedback generato in base ai criteri esplicitati, ma ne colgono con chiarezza anche i limiti pedagogici ed etici emersi nel confronto con il giudizio umano. L'IA è quindi percepita come strumento di supporto alla valutazione, non come sostituto del giudizio professionale del docente, che resta imprescindibile per cogliere complessità, singolarità

e dimensione relazionale dell'apprendimento. L'esperienza evidenzia la necessità di integrare percorsi sistematici di AI literacy nella formazione iniziale e in servizio. Una competenza critica nell'uso dell'IA non riguarda solo la padronanza tecnica, ma implica la comprensione delle logiche di funzionamento, delle condizioni di utilizzo e dei limiti, al fine di salvaguardare il senso educativo della valutazione. In questa prospettiva, l'IA può diventare un alleato per rendere più trasparente e consapevole l'agire valutativo, purché subordinata a una regia pedagogica esplicita e responsabile. Un ampliamento del campione e una maggiore diversificazione dei contesti consentirebbero di approfondire come percezioni e pratiche legate all'IA varino in base ad area disciplinare, esperienza professionale e concezioni di valutazione. Resta centrale la necessità di ripensare la valutazione nella scuola contemporanea come pratica formativa e relazionale, capace di resistere a riduzioni tecnicistiche mentre si esplorano in modo critico le opportunità offerte dalle tecnologie emergenti.

Riferimenti bibliografici

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). *Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters*. arXiv.

Brembilla, M. (2025). Embodied cognition e sfide educative: Per una pratica creativa e trasformativa. *Journal of Inclusive Methodology and Technology in Learning and Teaching, 5*(2).

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). *The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges*. arXiv.

Cohn, C., Hutchins, N., Le, T., & Biswas, G. (2024). *A chain-of-thought prompting approach with LLMs for evaluating students' formative assessment responses in science*. arXiv.

Creati, P., & Cuccaro, A. (2025). Special pedagogy and artificial intelligence: A digital humanism for inclusive education. *Italian Journal of Special Education for Inclusion, 13*(1), 202-213.

Ellerani, P., & Ferrari, L. (2024). The contribution of generative AI ecosystems in micro-instructional design: Opportunities and limitations. *Formazione & insegnamento, 22*(1), 117-124.

Fiorucci, A., & Bevilacqua, A. (2024). Il dibattito scientifico sull'Intelligenza Artificiale in ambito educativo: Una scoping review sugli approcci e sulle tendenze della ricerca pedagogica in Italia. *Education Sciences & Society, 2*, 416-433.

Fiorucci, A., & Bevilacqua, A. (2025). Rethinking artificial intelligence for inclusion and disability support: Trajectories and trends of Italian pedagogical research. *Journal of Inclusive Methodology and Technology in Learning and Teaching, 5*(1).

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences, 11*(12), 5467.

Mele, L. M., & Gentile, M. R. (2023). The enhancement of assessment through artificial intelligence in the context of the onlife ecosystem [La valorizzazione dell'assessment tramite l'Intelligenza Artificiale nel contesto dell'ecosistema onlife]. *QTimes – Webmagazine, 15*(4).

Panciroli, C., Rivoltella, P. C., Gabbrielli, M., & Zawacki-Richter, O. (2020). Artificial intelligence and education: New research perspectives [Intelligenza artificiale e educazione: Nuove prospettive di ricerca]. *Form@re – Open Journal per la formazione in rete, 20*(3), 1-12.

Perla, L., & Vinci, V. (2024). Rethinking assessment in the digital era: Designing a pilot study on hybridization in higher education. *QWERTY – Open and Interdisciplinary Journal of Technology, Culture and Education, 19*(1), 33-51.

Pillera, G. C. (2023). A dialogo con ChatGPT su potenzialità e limiti dell'IA per la valutazione in educazione. *Pedagogia oggi, 21*(1), 301-315.

Qian, Y. (2025). *Pedagogical applications of generative AI in higher education: A systematic review of the field. TechTrends*. Advance online publication.

Ranieri, M. (2024). Intelligenza artificiale a scuola. Una lettura pedagogico-didattica delle sfide e delle opportunità. *Studi sulla Formazione, 62*(1), 123–135.

Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: A comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education*. Advance online publication.

Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21, Article 15.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education*, 16, Article 39.

II.20
RubricaBot: un supporto intelligente alla progettazione valutativa nella formazione dei docenti

Marilena di Padova
Università Telematica Pegaso, marilena.dipadova@unipegaso.it
Angelo Basta
Università E-campus
Andrea Tinterri
Università Telematica Pegaso
Anna Dipace[1]
Università Telematica Pegaso

Abstract

Lo studio analizza l'usabilità percepita di RubricaBot, chatbot con IA generativa che supporta i docenti nella costruzione di rubriche valutative, in una logica *human-in-the-loop*. La sperimentazione, nel percorso di specializzazione per il sostegno dell'Università Telematica Pegaso, ha coinvolto 150 docenti di primaria; 119 hanno utilizzato il chatbot e compilato il questionario. Il disegno *mixed-methods* integra il Chatbot Usability Questionnaire (CUQ), tradotto e adattato in italiano, con domande aperte analizzate tematicamente. I risultati indicano alta accettabilità (CUQ medio=82,3; DS=13,6). Le risposte qualitative evidenziano rapidità, chiarezza e riduzione del carico cognitivo; criticità riguardano qualità del prompt, possibili effetti di standardizzazione e vincoli tecnici. Nel complesso, RubricaBot appare un supporto utile, da validare in contesti e discipline differenti.

Parole chiave: valutazione; chatbot; intelligenza artificiale generativa; rubriche.

The study examines the perceived usability of RubricaBot, a generative-AI chatbot that supports teachers in designing assessment rubrics within a *human-in-the-loop* approach. The pilot was conducted in the specialisation programme for

1 Il contributo è il risultato di un lavoro condiviso. Ad ogni modo, ai fini dell'attribuzione Marilena di Padova ha scritto i paragrafi 3 e 4; Angelo Basta il paragrafo 1; Andrea Tinterri il paragrafo 2; Anna Dipace i paragrafi Introduzione e 5.

support teachers at Pegaso University and involved 150 primary school teachers; 119 participants used the chatbot and completed the questionnaire. The *mixed-methods* design combines the Chatbot Usability Questionnaire (CUQ), translated and adapted into Italian, with open-ended questions analysed through thematic analysis. Results show high acceptability (mean CUQ = 82.3; SD = 13.6). Qualitative responses highlight speed, clarity, and reduced cognitive load, while reported limitations concern prompt quality, potential standardisation effects, and technical constraints. Overall, RubricaBot appears to be a useful support tool that warrants further validation across different contexts and subject areas.

Keywords: assessment; chatbot; Generative Artificial Intelligence; rubrics.

Introduzione

Negli ultimi anni, i chatbot basati su intelligenza artificiale hanno progressivamente assunto un ruolo sempre più rilevante nei contesti educativi, configurandosi come strumenti in grado di supportare processi di apprendimento, progettazione didattica e riflessione professionale (Davar, Dewan & Zhang, 2025; Gökçearslan, Tosun & Erdemir, 2024; Hwang & Chang, 2023). I chatbot di nuova generazione si caratterizzano per una crescente capacità di interazione linguistica, adattamento contestuale e generazione di contenuti, aprendo scenari inediti per l'integrazione dell'intelligenza artificiale nella pratica didattica quotidiana (Maher, 2026). In questo quadro, l'attenzione della ricerca si è progressivamente spostata dalla mera efficacia tecnica di tali strumenti alla qualità dell'esperienza d'uso, al loro impatto sui processi cognitivi degli utenti e alle condizioni che ne rendono l'impiego pedagogicamente significativo (Badali et al., 2026; Badger, López & Nissen, 2026).

Un'evoluzione particolarmente rilevante è rappresentata dalla diffusione dei chatbot personalizzati, ossia sistemi configurati per rispondere a bisogni specifici di contesto, di disciplina o di funzione didattica (Sinha et al., 2019). A differenza dei chatbot generalisti, questi strumenti consentono di modellare il comportamento conversazionale in funzione di obiettivi educativi espliciti, vincoli metodologici e pratiche professionali consolidate, trasformando l'interazione in un dispositivo di mediazione piuttosto che in una semplice interrogazione informativa (Kuhail et al., 2023; Gonda & Chu, 2019). In ambito didattico, tale personalizzazione solleva questioni centrali non solo in termini di potenzialità, ma anche di sostenibilità pedagogica: l'efficacia di un chatbot non dipende esclusivamente dalla qualità degli output generati, bensì dalla sua capacità di suppor-

tare l'azione riflessiva del docente di preservare l'autonomia decisionale dell'utente e di ridurre il carico cognitivo e il *time-consuming* (Singh et al., 2025). In questa prospettiva, diventa cruciale interrogarsi su come i chatbot personalizzati vengano effettivamente percepiti e utilizzati dagli insegnanti, quali forme di usabilità emergano dall'interazione e in che misura tali strumenti riescano a configurarsi come alleati cognitivi all'interno delle pratiche educative. È all'interno di questo quadro che si colloca il presente studio, orientato a esplorare l'esperienza d'uso di un chatbot personalizzato per la didattica attraverso un'analisi integrata quantitativa e qualitativa.

1. RubricaBot: design di un chatbot personalizzato

Nel quadro delineato, RubricaBot è stato progettato come chatbot didattico personalizzato finalizzato a supportare i docenti nella progettazione e costruzione di rubriche valutative. La costruzione di rubriche può essere un processo molto complesso, che richiede una profonda comprensione della disciplina da parte del progettista nonché tempo (Chen, Chen & Lin, 2020). Realizzato su una piattaforma di intelligenza artificiale generativa (ChatGPT), RubricaBot è stato configurato per operare come dispositivo orientato a un compito didattico specifico, integrando criteri metodologici, vincoli procedurali e un linguaggio coerente con le pratiche professionali della valutazione educativa. Rappresenta un artefatto di mediazione cognitiva, capace di accompagnare l'utente lungo le diverse fasi della progettazione valutativa, riducendo il carico cognitivo iniziale e rendendo esplicita la struttura del compito, anche in ottica inclusiva (CAST, 2024; Montanari & Sibi, 2024). La progettazione del chatbot ha perseguito l'obiettivo di offrire un supporto strutturato e guidato alla definizione di dimensioni, indicatori e livelli di prestazione per la definizione condivisa di una rubrica di competenza, senza sostituirsi alle decisioni del docente, ma piuttosto sostenendone il processo riflessivo e organizzativo, utilizzando il modello delineato, tra gli altri, da Castoldi (2019). In questa prospettiva, l'impostazione di RubricaBot assume la rubrica non come semplice griglia descrittiva, ma come dispositivo di mediazione che rende espliciti criteri e livelli, sostiene l'allineamento tra compito, obiettivi e giudizio e facilita la trasformazione di intenzioni valutative in descrittori osservabili. Il prompt design è stato quindi orientato a guidare l'utente nella definizione progressiva di dimensioni, indicatori e livelli di prestazione, sollecitando coerenza interna, distinguibilità dei livelli e chiarezza linguistica. Tale configurazione intende preservare la centralità decisionale del docente, riconoscendo che la qualità di una rubrica di-

pende dalla sua contestualizzazione didattica e dalla negoziazione professionale dei criteri.

La realizzazione del chatbot si è fondata su una fase di prompt design strutturato (Fig. 1), volta a definire in modo esplicito il ruolo del sistema, i confini del dominio di intervento e i criteri metodologici che regolano la generazione delle risposte (Singiri et al., 2026).

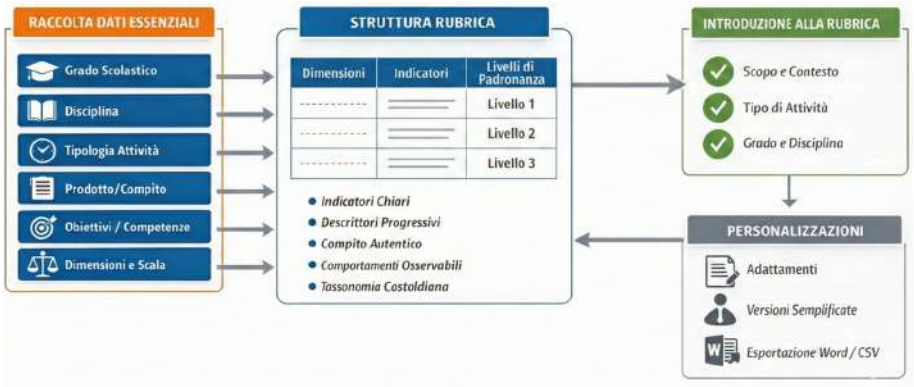


Fig. 1 – Prompt design di RubricaBot

Dal punto di vista dell'architettura conversazionale, il chatbot è stato progettato per favorire un'interazione di tipo incrementale e dialogico, in cui ogni risposta è condizionata dalle informazioni precedentemente fornite dall'utente. Tale impostazione consente al sistema di adattare dinamicamente l'output, mantenendo coerenza interna e continuità semantica nel corso della conversazione (Kuhail et al., 2023; Debnath & Agarwal, 2020). Nel suo complesso, il design di RubricaBot si colloca all'incrocio tra architettura tecnica e intenzionalità pedagogica (Darshan et al., 2025) e questa impostazione costituisce il presupposto per l'analisi empirica dell'esperienza d'uso presentata nel presente studio. A partire da questa impostazione, il presente studio si propone di analizzare in che modo RubricaBot venga percepito e utilizzato dagli utenti, con particolare attenzione alla sua usabilità e alle rappresentazioni cognitive e professionali che emergono dall'interazione.

2. Metodi e strumenti

RubricaBot è stato sperimentato all'interno di una lezione laboratoriale inserita nel percorso di specializzazione per il sostegno (D.M. n. 75 e D.I. n. 77 del 24 aprile 2025) dell'Università Telematica Pegaso, rivolta a un gruppo complessivo di 150 docenti della scuola primaria. La lezione aveva come obiettivo formativo l'approfondimento delle rubriche di valutazione, con particolare attenzione alla loro funzione pedagogica, alla struttura interna e alle fasi di costruzione. Nella prima parte dell'attività, la lezione ha previsto una spiegazione teorico-operativa delle rubriche valutative, affrontando i concetti di dimensione, indicatori e livelli di prestazione, nonché le principali criticità riscontrate nella pratica scolastica. Successivamente, l'attività laboratoriale è stata integrata con l'utilizzo di RubricaBot come strumento di supporto alla progettazione. Il chatbot è stato presentato come risorsa opzionale e complementare al lavoro del docente, consistente nella progettazione di una rubrica di prestazione. Al termine dell'attività laboratoriale, ai partecipanti è stato somministrato un questionario per raccogliere una valutazione dell'esperienza d'uso. Su un totale di 150 docenti coinvolti nella lezione, 119 partecipanti hanno effettivamente utilizzato il chatbot e compilato il questionario, costituendo il campione di riferimento per l'analisi quantitativa e qualitativa presentata nel presente studio.

La valutazione dell'usabilità del chatbot è stata condotta mediante il *Chatbot Usability Questionnaire (CUQ)* (Holmes et al., 2019), strumento standardizzato sviluppato specificamente per l'analisi dell'esperienza d'uso di sistemi conversazionali, Il CUQ si colloca nel solco della tradizione degli strumenti di valutazione soggettiva dell'usabilità, assumendo come riferimento teorico il *System Usability Scale (SUS)* (Brooke, 2013), dal quale riprende l'impostazione basata su item bilanciati e su una restituzione sintetica del giudizio complessivo. Tuttavia, a differenza del SUS destinato all'utilizzo più generico di sistemi informatici computerizzati, il CUQ è progettato per intercettare le peculiarità dell'interazione dialogica uomo-chatbot. Nel caso specifico, si tratta di un adattamento in lingua italiana del questionario, tradotto appositamente per questo studio. Il CUQ è stato tradotto e adattato in italiano mediante doppia traduzione indipendente (due ricercatori), armonizzazione, revisione linguistica da parte di un madrelingua e controllo interno finale, mantenendo invariata la struttura originale (16 item bilanciati, scala Likert a 5 punti, *reverse-coding* degli item negativi e scoring normalizzato 0-100). La versione italiana è stata somministrata una prima volta in un laboratorio per docenti del percorso di specializzazione per il sostegno (agosto 2025; N=87), mostrando una buona coerenza interna (α=0,835), e successivamente nel campione dello studio (N=119), in

cui l'affidabilità interna risulta elevata (α=0,897). Le evidenze di validità considerate in questa fase riguardano soprattutto la validità di contenuto e di facciata legate alla procedura di adattamento linguistico; ulteriori studi potranno approfondire la validità di costrutto e la stabilità temporale in campioni più ampi e diversificati.

Il questionario è composto da sedici affermazioni, di cui otto formulate in senso positivo e otto in senso negativo, valutate attraverso una scala Likert a cinque punti. Il punteggio complessivo del CUQ è ottenuto mediante una procedura di normalizzazione che consente di esprimere l'usabilità percepita su una scala continua da 0 a 100 (Fig. 2), rendendo lo strumento adatto sia a confronti intra-campione sia a confronti tra sistemi differenti.

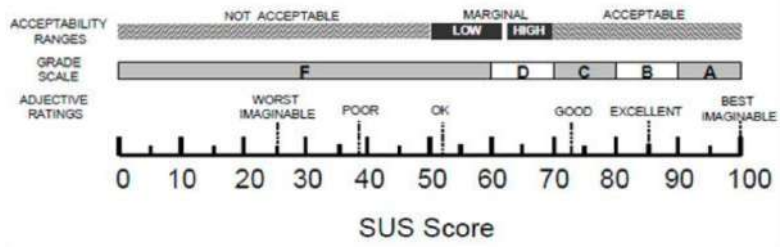


Fig. 2 – Scala di valutazione SUS

Per l'analisi dei dati è stato utilizzato lo strumento di calcolo CUQ (Calculation Tool) rappresentato da un foglio di calcolo Microsoft Excel che può essere utilizzato come mezzo rapido per calcolare i punteggi CUQ singoli e complessivi.

In coerenza con un approccio metodologico di tipo *mixed-methods*, il questionario standardizzato è stato integrato da una serie di domande aperte, progettate per ampliare e approfondire l'interpretazione del dato quantitativo. In particolare, le domande aperte hanno avuto la funzione di esplorare: i punti di forza percepiti nell'utilizzo del chatbot; le eventuali criticità o limiti riscontrati; i suggerimenti di miglioramento; le rappresentazioni metaforiche associate all'esperienza d'uso (queste ultime non incluse nel presente studio). Tale integrazione risponde alla consapevolezza che le misure standardizzate di usabilità, pur offrendo una sintesi affidabile e comparabile, non sono di per sé diagnostiche e non consentono di cogliere la complessità del vissuto soggettivo, delle aspettative e delle attribuzioni di significato da parte degli utenti.

L'analisi qualitativa è stata condotta seguendo l'approccio dell'analisi tematica delineato da Braun e Clarke (2006). Il processo di codifica ha integrato tecniche

di *open e axial coding* (Saldaña, 2021; Strauss & Corbin, 1998) per strutturare le categorie, mentre la validazione dei temi è avvenuta verificando i criteri di omogeneità interna ed eterogeneità esterna (Patton, 2015).

3. Analisi quantitativa dei dati

L'analisi quantitativa dell'usabilità percepita riporta un punteggio medio complessivo risulta pari a 82,3 con una deviazione standard di 13,6, indicando un livello elevato di usabilità percepita accompagnato da una variabilità interindividuale moderata. Il range dei punteggi osservati è ampio, con un valore minimo di 39,1 e un valore massimo di 100,0. La mediana, pari a 84,4, risulta superiore alla media aritmetica, suggerendo una distribuzione leggermente asimmetrica, con una maggiore concentrazione dei punteggi nella fascia alta della scala e una coda di valori più bassi.

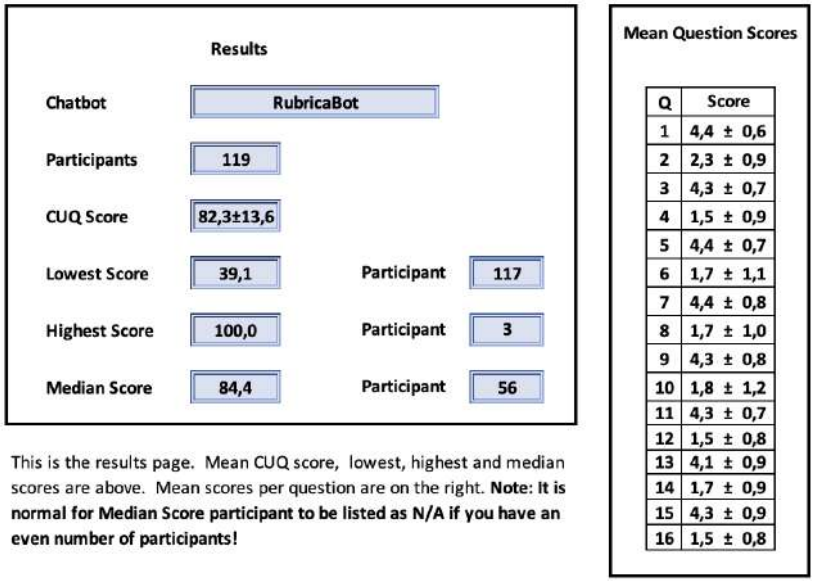


Fig. 3 – Risultati del CUQ

In riferimento ai criteri interpretativi, mutuati dalla letteratura sul *System Usability Scale (SUS),* un punteggio superiore a 80 colloca il sistema nella fascia di alta accettabilità, associata a valutazioni qualitative comprese tra *good* ed *excellent*.

L'analisi descrittiva dei punteggi medi per singolo item consente di esaminare in modo approfondito la struttura interna delle valutazioni espresse, mettendo in relazione i valori numerici con le specifiche dimensioni di usabilità indagate.

Gli item formulati positivamente presentano valori medi sistematicamente elevati e relativamente omogenei, collocandosi stabilmente nella porzione superiore della scala Likert. In particolare, gli item Q1, Q5 e Q7 (tutti con media 4,4, DS compresa tra 0,6 e 0,8), riferiti rispettivamente alla percezione di semplicità d'uso, alla facilità di apprendimento dell'interazione e alla fluidità complessiva dell'esperienza, evidenziano una valutazione fortemente positiva e condivisa di questi aspetti fondamentali dell'usabilità. Valori analoghi si riscontrano negli item Q3, Q9, Q11 e Q15 (medie comprese tra 4,3 e 4,4), che indagano dimensioni quali la chiarezza delle risposte fornite, la capacità del chatbot di comprendere correttamente gli input dell'utente, la coerenza dell'interazione e la percezione di utilità del sistema nel contesto d'uso.

In modo complementare, gli item formulati negativamente presentano valori medi bassi, prevalentemente concentrati nella parte inferiore della scala Likert. Gli item Q4, Q12 e Q16 (tutti con media 1,5 e deviazioni standard comprese tra 0,8 e 0,9), riferiti rispettivamente alla percezione di confusione, alla sensazione di inutilità del sistema e alla difficoltà complessiva di utilizzo, mostrano un diffuso e marcato disaccordo, indicando che tali caratteristiche negative non vengono riconosciute come proprie dell'esperienza vissuta dalla maggioranza dei partecipanti.

Valori analoghi si osservano per gli item Q6, Q8 e Q14 (medie pari a 1,7, con deviazioni standard comprese tra 0,9 e 1,1), che indagano aspetti legati alla rigidità dell'interazione, alla frustrazione potenziale durante l'uso e alla necessità di supporto esterno per utilizzare il chatbot. In questi casi, la dispersione leggermente più elevata delle risposte suggerisce una variabilità interindividuale più pronunciata, pur in presenza di una valutazione complessivamente negativa delle criticità ipotizzate.

L'item Q10 "Il chatbot non ha riconosciuto molti dei miei input" (1,8 ± 1,2), relativo a una dimensione particolarmente sensibile dell'usabilità, presenta la deviazione standard più elevata tra gli item negativi, indicando una maggiore eterogeneità nelle risposte e suggerendo che tale aspetto venga vissuto in modo differenziato dai partecipanti. L'item Q2 "Il chatbot sembrava troppo robotico" (2,3 ± 0,9), pur rimanendo al di sotto del punto neutro della scala, si colloca su valori relativamente più alti rispetto agli altri item negativi; esso esplora una dimensione di complessità percepita dell'interazione, e il suo posizionamento intermedio indica una valutazione meno polarizzata, con una quota di partecipanti che esprime giudizi più cauti o ambivalenti.

4. Analisi qualitativa dei dati

I risultati dell'analisi tematica, organizzati per ciascuna domanda aperta, delineano un profilo qualitativo articolato. In relazione ai punti di forza (112 risposte), emerge anzitutto il Tema A1 – Facilità d'uso e intuibilità (29%), in cui il chatbot è descritto come "accessibile", "immediato" e "semplice da usare"; esempi ricorrenti includono formulazioni quali "*…ben strutturato… di facile accesso*" e "*velocità ed efficienza, facile da usare*". Segue il Tema A2 – Rapidità e risparmio di tempo (26%), incentrato sulla compressione dei tempi di progettazione/valutazione e sull'efficienza operativa, spesso espressa con enunciati come "*snelliamo i tempi*" o con enfatizzazioni valutative quali "*meravigliosamente rapido*". Il Tema A3 – Chiarezza e guida procedurale (20%) mette in evidenza la funzione di orientamento del chatbot nei compiti complessi: RubricaBot viene rappresentato come uno strumento che "mette ordine" e aiuta a sequenziare l'azione, con risposte come "*Chiarezza… linguaggio appropriato*". Il Tema A4 – Qualità/pertinenza delle risposte (16%) riguarda la percezione di utilità informativa e coerenza dell'output ("*coerenza nelle risposte*"), mentre il Tema A5 – Supporto specifico a rubriche/valutazione (13%) valorizza la funzione core legata alla costruzione di rubriche, spesso evidenziata da commenti del tipo "*Con poche indicazioni… una rubrica dettagliata*" o "*aiuta chi non ha mai scritto rubriche*".

Per quanto concerne i punti di debolezza (104 risposte), il dato qualitativamente più informativo risiede nella tipologia di criticità che emergono nonostante l'alto gradimento complessivo. Il Tema B0 – Nessuna criticità rilevata (23%) raccoglie risposte del tipo "nessuna/nessuno", indicativo di accettazione e soddisfazione globale. Il Tema B1 – Necessità di prompt/istruzioni precise (13%) riguarda difficoltà legate alla messa a fuoco della richiesta e alla comprensione dello scopo operativo ("*Poca spiegazione su cosa potessi fare*"; "*Necessità di indicazioni precise*"). Il Tema B2 – Personalizzazione/contestualizzazione limitata (9%) esprime il timore di generalizzazione e il bisogno di rifinitura da parte del docente ("*…attenzione che non venga troppo generalizzato*"; "*in alcuni casi limitato*"). Il Tema B3 – Problemi tecnici o dipendenza dalla tecnologia (7%) include osservazioni su accesso/account e vincoli tecnici (es. difficoltà di login; "*Dipendenza dalla tecnologia*"). Il Tema B4 – Rischio di standardizzazione (3%) tematizza la possibilità che l'output renda le produzioni "tutte simili" ("*tutti ci ritroveremo con le stesse risposte*"), mentre il Tema B5 – Linguaggio robotico / supporto visivo (2% + 2%) segnala criticità minoritarie ma presenti relative al tono conversazionale e agli aspetti di visualizzazione.

In relazione ai suggerimenti di miglioramento (94 risposte), il Tema C0 – Nessuna modifica (49%) risulta particolarmente rilevante, poiché indica che

circa metà dei rispondenti non propone cambiamenti, in coerenza con una valutazione complessiva positiva. Tra coloro che formulano indicazioni, emerge C1 – Linguaggio più naturale/migliore (4%), C2 – Integrazioni/estensioni (4%) (con richieste orientate a una versione "pro" e alla capacità del sistema di apprendere dalle correzioni del docente), C3 – Più personalizzazione (3%), C4 – Più esempi/template (2%) e C5 – Migliorie grafiche/visual (1%).

5. Discussione e conclusioni

I risultati dello studio consentono di collocare l'esperienza d'uso di RubricaBot all'interno di una prospettiva che supera una concezione meramente strumentale dell'usabilità, restituendone una lettura di natura cognitiva, pedagogica e professionale. L'elevato punteggio CUQ, associato a una distribuzione fortemente polarizzata verso valori alti e a una dispersione complessivamente contenuta, indica una percezione condivisa di elevata usabilità. Tuttavia, è l'integrazione con i dati qualitativi a chiarire in modo più profondo il significato di tale valutazione: l'usabilità non emerge semplicemente come assenza di problemi tecnici, ma come qualità dell'interazione capace di sostenere il pensiero e l'azione del docente in un compito professionalmente rilevante.

In particolare, le risposte aperte mostrano come l'usabilità venga costruita dagli utenti prevalentemente in termini di usabilità cognitiva, ovvero come capacità del chatbot di ridurre il carico cognitivo iniziale, rendere esplicita la struttura del compito e supportare la sequenzialità delle decisioni progettuali. RubricaBot non viene percepito come un mero generatore automatico di contenuti, ma come un dispositivo che orienta, accompagna e organizza il ragionamento, soprattutto in un ambito – quello della costruzione delle rubriche valutative – spesso vissuto come complesso, *time-consuming* e cognitivamente impegnativo. Letti alla luce dei modelli teorici sulle rubriche valutative, i risultati indicano un esito pedagogicamente rilevante: l'elevata accettabilità e i temi qualitativi su chiarezza, guida procedurale e riduzione del carico cognitivo suggeriscono che RubricaBot sostenga i passaggi più onerosi della costruzione delle rubriche. L'usabilità percepita, dunque, non è solo un dato tecnico, ma segnala un supporto alla progettazione valutativa in senso formativo. Al contempo, le criticità (necessità di prompt precisi, personalizzazione limitata e rischio di standardizzazione) confermano che la rubrica non può essere trattata come template indifferenziato: richiede contestualizzazione e validazione professionale, rendendo decisivo l'approccio *human-in-the-loop*. In chiave UDL, la rubrica può funzionare come dispositivo di accessibilità valutativa solo se criteri e descrittori

restano chiari e consentono evidenze plurime; altrimenti, descrittori generici o standardizzati rischiano di indebolire la personalizzazione e l'equità.

Le criticità segnalate dai docenti suggeriscono che l'Intelligenza Artificiale non deve operare in una logica di sostituzione funzionale, bensì di potenziamento delle capacità umane. Se l'obiettivo fosse la sostituzione, il docente accetterebbe passivamente l'automazione; al contrario, la ricerca attiva di personalizzazione e la cura nel definire gli input in ottica di prompt engineering, testimoniano un approccio orientato all'*Intelligence Augmentation* (IA) (Zhou et al. 2023). Il valore aggiunto non risiede nell'automatizzare compiti cognitivi per escludere l'uomo, ma nell'utilizzare le macchine per estendere la portata e la precisione del giudizio umano (Davenport & Kirby, 2016)

RubricaBot, infatti, si configura quindi come uno strumento 'non autosufficiente' per la progettazione, la cui efficacia dipende strettamente dalla presenza di un *Human-in-the-loo*p (Qiu et al., 2025). Questo concetto ribadisce che il docente non è un passivo fruitore, ma mantiene il controllo critico e valoriale sul processo. Gli utenti vedono nell'AI un alleato per l'efficienza operativa, ma rivendicano il ruolo centrale nella validazione pedagogica, rifiutando quella che Luckin et al. (2016) definiscono una delega cieca all'algoritmo in favore di una 'collaborazione interattiva'. I risultati suggeriscono, infatti, che RubricaBot possa agire come *scaffold* cognitivo e procedurale nella fase di avvio della rubrica, soprattutto quando il docente deve trasformare obiettivi/competenze in dimensioni osservabili, indicatori e livelli di prestazione. In termini operativi, l'utilità maggiore non risiede nella produzione automatica della rubrica, ma nella possibilità di rendere più fluido e meno oneroso il passaggio dal quadro intenzionale alla formalizzazione valutativa, mantenendo il docente in una posizione di controllo e validazione (human-in-the-loop). In una prospettiva di adozione realistica, RubricaBot può supportare almeno tre funzioni: (a) avvio e strutturazione della rubrica, offrendo una prima architettura coerente con il compito; (b) rifinitura e miglioramento iterativo, attraverso cicli brevi di revisione guidata; (c) allineamento e trasparenza, rendendo più esplicite le scelte valutative e favorendo una comunicazione chiara verso studenti e famiglie. Perché tali ricadute si traducano in pratiche effettive, lo strumento dovrebbe essere integrato in routine professionali riconoscibili (es. dipartimenti, team docenti, comunità di pratiche), accompagnato da micro-competenza di *prompt engineering* (come formulare evidenze osservabili, vincolare il contesto, chiedere esempi e contro-esempi) e da procedure di controllo qualità.

Accanto a questi elementi di forza, lo studio presenta alcuni limiti che è opportuno esplicitare. In primo luogo, il campione è costituito esclusivamente da docenti della scuola primaria inseriti in un percorso di specializzazione per il so-

stegno, potenzialmente più orientati a progettazione inclusiva e strumenti di personalizzazione, con possibili effetti di contesto formativo e di novità. La partecipazione effettiva (119/150) può inoltre indicare un bias di autoselezione, privilegiando docenti più motivati o digitalmente competenti. Infine, l'assenza di dati su anzianità, familiarità con le rubriche e competenze digitali limita analisi differenziali e, quindi, la generalizzabilità dei risultati. In secondo luogo, la sperimentazione si è svolta all'interno di una singola esperienza laboratoriale, circoscritta nel tempo, e non consente di osservare l'evoluzione dell'esperienza d'uso nel medio-lungo periodo o l'impatto del chatbot sulle pratiche didattiche consolidate. Inoltre, i dati raccolti si basano su misure di percezione soggettiva dell'usabilità e non includono indicatori oggettivi di performance o analisi longitudinali dell'effettiva qualità delle rubriche prodotte. Infine, si ricorda che il questionario utilizzato è un adattamento in lingua italiana del CUQ che è in lingua inglese.

Nonostante tali limiti, i risultati offrono indicazioni rilevanti per la progettazione e l'adozione di chatbot personalizzati in ambito educativo, suggerendo la necessità di considerare l'intelligenza artificiale come alleato cognitivo e non come dispositivo sostitutivo. Studi futuri potranno ampliare il campione, diversificare i contesti di applicazione e integrare ulteriori livelli di analisi, al fine di approfondire il ruolo dei chatbot come strumenti di mediazione nei processi di insegnamento, valutazione e formazione dei docenti

Riferimenti bibliografici

Badali, M., Shahalizadeh, M., Alimirzaie, M., Noroozi, O., & Kazem Banihashem, S. (2026). Bridging AI chatbot use and critical thinking: the mediating role of AI literacy. *Interactive Learning Environments*, 1-13.

Badger, M., López, N., & Nissen, C. F. R. (2026). The Impact of GenAI Chatbots on Student Learning in Higher Education: A Literature Review. *International Journal of Technology in Education*, *9*(1), 43-69.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.

Brooke, J. (2013). SUS: a retrospective. *Journal of usability studies*, *8*(2).

CAST (2024), *Universal Design for Learning Guidelines version 3.0*, Lynnfield, CAST, traduzione e adattamento della versione italiana a cura di Lia Daniela Sasanelli.

Castoldi, M. (2019). *Rubriche valutative: guidare l'espressione del giudizio*. UTET università.

Chen, L., Chen, P. & Lin, Z. (2020). Artificial Intelligence in education: a review. *IEEE Access* (8). pp. 75264–75278

Darshan, P., Sanghavi, E., Chennakeshava, G., Kodabagi, M. M., Ali, B. B. F., & Ga-

yathri, H. P. (2025, April). AI Chatbot for Educational Institution: A Case Study. In *2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1-5). IEEE.

Davar, N. F., Dewan, M. A. A., & Zhang, X. (2025). AI chatbots in education: challenges and opportunities. *Information*, *16*(3), 235.

Davenport, T. H., & Kirby, J. (2016). *Only humans need apply: Winners and losers in the age of smart machines.* Harper Business.

Debnath, B., & Agarwal, A. (2020). A framework to implement AI-integrated chatbot in educational institutes. *J. Stud. Res*, *1*, 16.

Gökçearslan, S., Tosun, C., & Erdemir, Z. G. (2024). Benefits, challenges, and methods of artificial intelligence (AI) chatbots in education: A systematic literature review. *International Journal of Technology in Education*, *7*(1), 19-39.

Gonda, D.E., Chu, B.: Chatbot as a learning resource? Creating conversational bots as a supplement for teaching assistant training course. In: *IEEE International Conference on Engineering, Technology and Education* (TALE) (2019). https://doi.org/10.1109/tale48000.2019.9225974

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & Mctear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?. In *Proceedings of the 31st European Conference on Cognitive Ergonomics* (ECCE 2019), Maurice Mulvenna and Raymond Bond (Eds.). ACM, New York, NY, USA, 207-214.

Hwang, G. J., & Chang, C. Y. (2023). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, *31*(7), 4099-4112.

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973-1018.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence Unleashed: An argument for AI in Education*. Pearson.

Maher, D. (2026). Chatbots in schools: Opportunities and considerations. In *Social robots and artificial intelligence in education: Integrating AI in K-12 and higher education* (pp. 303-326). Cham: Springer Nature Switzerland.

Montanari, M., & Sibi, P. (2024). Intelligenza artificiale e educazione. Nuove sfide per la formazione degli insegnanti inclusivi. *Nuova secondaria*, *41*(8).

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (4th ed.). Sage.

Qiu, W., Thway, M., Lai, J. W., & Lim, F. S. (2025, March). GenAI for teaching and learning: A Human-in-the-loop Approach. In *Proceedings of the Companion Proceedings 15th International Conference on Learning Analytics & Knowledge* (LAK25), Dublin, Ireland (pp. 3-7).

Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Sage.

S., Sunil Kumar, M., Yethish, Y., Neelima, P., Ganesh, D., & Sushama, C. (2024, August). Building a Custom Chatbot for the Educational Domain. In *International Conference on Advances in Computer Engineering and Communication Systems* (pp. 71-79). Singapore: Springer Nature Singapore.

Singh, H., Natrajan, N. S., Sanjeev, R., & Singh, A. (2025). Transformative Pedagogy: ChatGPT as a Catalyst for Educational Innovation. In *The ChatGPT Revolution* (pp. 153-182). Emerald Publishing Limited.

Singiri, S., Sunil Kumar, M., Yethish, Y., Neelima, P., Ganesh, D., Sushama, C. (2026). Building a Custom Chatbot for the Educational Domain. In: Nagini, S., Govardhan, A., Baby, V., Kountcheva, R., Raju, K.S. (eds) *Proceedings of Fifth International Conference on Advances in Computer Engineering and Communication Systems*. ICACECS 2024. Smart Innovation, Systems and Technologies, vol 117. Springer, Singapore. https://doi.org/10.1007/978-981-96-4410-0_7

Sinha, S., Basak, S., Dey, Y., & Mondal, A. (2019). An educational chatbot for answering queries. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018* (pp. 55-60). Singapore: Springer Singapore.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.

Zhou, L., Rudin, C., Gombolay, M., Spohrer, J., Zhou, M., & Paul, S. (2023). From artificial intelligence (AI) to intelligence augmentation (IA): Design principles, potential risks, and emerging issues. *AIS Transactions on Human-Computer Interaction*, 15(1), 111-135.

II.21
Progettare la valutazione con l'algoritmo. Uno studio sull'impatto dell'IA sulla competenza valutativa dei docenti pre-service

Viviana Vinci[1]
Università degli Studi di Foggia, viviana.vinci@unifg.it
Pierangelo Berardi[2]
Università degli Studi di Foggia, pierangelo.berardi@unifg.it

Abstract

Il presente lavoro indaga il ruolo dell'Intelligenza Artificiale Generativa (GenAI) come partner riflessivo nella competenza valutativa dei docenti *pre-service*. Attraverso un disegno *mixed-methods* (QUAN-qual) su 191 studenti di Scienze della Formazione Primaria, si analizza la co-progettazione di rubriche assistita dall'IA. I risultati mostrano che l'IA funge da *scaffolding* efficace (M = 8,30), senza sostituire l'intenzionalità pedagogica. La *cluster analysis* identifica tre profili (Sperimentatori, Mediatori, Prudenti), legando l'integrazione alla *teacher agency* individuale. Emerge una forte focalizzazione verso strategie inclusive per i Bisogni Educativi Speciali. In conclusione, la ricerca suggerisce un'evoluzione verso una *AI Literacy* ermeneutica, capace di mediare criticamente tra suggerimenti algoritmici e giudizio professionale situato.

Parole chiave: Intelligenza Artificiale; Competenza valutativa; Formazione docenti; Agency docente; Progettazione della valutazione.

1 Professore ordinario in Pedagogia sperimentale presso l'Università di Foggia. Componente del Gruppo di lavoro ANVUR *Riconoscimento e valorizzazione delle competenze didattiche della docenza universitaria* e della *Commissione per la redazione delle Indicazioni nazionali per il curricolo scolastico*. Referente eTwinning for Future Teachers ITE. Coordinatrice del tirocinio. Vincitrice del Premio SIRD 2021 e del Premio SIPED 2014.

2 Dottorando del corso LEDIEL presso l'Università degli Studi di Bari "Aldo Moro", docente a contratto presso il Dipartimento di Studi Umanistici. Lettere, Beni Culturali e Scienze della Formazione dell'Università degli Studi di Foggia, vincitore del premio Junior Researcher Award 2025 Visual Education Astera Conference.

This paper investigates the role of Generative Artificial Intelligence (GenAI) as a reflective partner in developing pre-service teachers' assessment competence. Using a mixed-methods design (QUAN-qual) on 191 students, AI-assisted rubric co-design is examined. Results reveal that AI serves as effective *scaffolding* (M = 8.30) without replacing pedagogical intentionality. Cluster analysis identifies three profiles (Experimenters, Mediators, Prudent users), linking technology integration to individual teacher agency. A significant focus on inclusive strategies for Special Educational Needs emerged. In conclusion, the research suggests an evolution towards a hermeneutic *AI Literacy*, capable of critically mediating between algorithmic suggestions and situated professional judgment.

Keywords: Artificial Intelligence; Assessment Literacy; Teacher Education; Teacher Agency; Assessment Design.

1. Verso un'architettura ibrida

L'ascesa dell'Intelligenza Artificiale Generativa (GenAI) sta determinando una riconfigurazione radicale dei paradigmi della professionalità docente, agendo come una forza trasformativa che investe prioritariamente la competenza valutativa e le architetture della progettazione didattica (Selwyn, 2019; Perla, Agrati, & Vinci, 2021). Mentre la letteratura prevalente si è focalizzata sull'IA come strumento finalizzato all'automatizzazione dei processi valutativi rivolti agli studenti (Holmes et al., 2019; Thanh et al., 2023), rimane ancora parzialmente inesplorato il potenziale di tali tecnologie nell'agire come catalizzatore per lo sviluppo della competenza valutativa del docente stesso. In questo scenario, il presente studio concettualizza l'IA come un "co-designer" euristico e un partner dialogico (Vinci & Berardi, 2025), capace di operazionalizzare il *Pedagogical Content Knowledge* (Shulman, 1987) attraverso una negoziazione continua tra la conoscenza formalizzata dell'algoritmo e il giudizio professionale situato dell'insegnante. Tale interazione si inserisce in una prospettiva di pedagogia post-digitale, in cui la relazione uomo-macchina non è più intesa come strumentale, ma come una co-abitazione in ecosistemi educativi ibridi (Fawns, 2022).

All'interno di questa architettura, il feedback generato dall'IA assume la funzione di uno *scaffolding* digitale (Wood, Bruner & Ross, 1976; Williams, 2025) in grado di innescare una "dissonanza produttiva" nei modelli mentali del progettista. Tale stimolo metacognitivo obbliga il docente *pre-service* a esplicitare e ristrutturare le proprie concezioni didattiche, favorendo l'evoluzione della *assessment literacy* (Stiggins, 1991) da una dimensione puramente tecnica a una riflessiva e critica. Affinché questa partnership euristica risulti efficace, l'output

algoritmico deve manifestare elevati standard di qualità, chiarezza, specificità e orientamento all'azione (Striano, Melacarne & Oliverio, 2018), che sono prerequisiti essenziali per la percezione di utilità e l'accettazione tecnologica (Davis, 1989). Tuttavia, il solo possesso di un feedback di qualità non garantisce l'apprendimento professionale se non è supportato da una matura *feedback literacy*, ovvero dalla capacità del docente di dare senso alle informazioni ricevute e utilizzarle per migliorare attivamente il proprio design (Carless & Winstone, 2020; Boud & Molloy, 2013).

2. Obiettivi dell'indagine e prospettive analitiche

Sulla base del quadro teorico delineato e delle sfide poste dall'integrazione dell'IA Generativa nella formazione iniziale, il presente lavoro si è posto l'obiettivo di rispondere a quattro domande di ricerca fondamentali, capaci di orientare l'approccio *mixed-methods* adottato. In primo luogo, l'indagine ha inteso quantificare il grado di utilità percepita dai futuri docenti nell'uso dell'IA come *scaffolding* digitale per la progettazione valutativa, analizzando al contempo l'entità dell'impatto trasformativo che tale strumento esercita sulle proposte progettuali originali. In secondo luogo, lo studio ha cercato di verificare se l'esperienza professionale pregressa possa costituire una variabile discriminante nella percezione dello strumento e quale sia il nesso statistico tra l'*agency* docente dei partecipanti e la loro effettiva propensione alla negoziazione critica con l'output algoritmico. Una terza direttrice di ricerca ha riguardato l'esplicitazione delle strategie di mediazione pedagogica adottate, con un focus specifico sulla capacità dei docenti in formazione di oggettivare i propri criteri valutativi impliciti attraverso il dialogo con la macchina e di declinare i suggerimenti dell'IA in chiave inclusiva per rispondere ai Bisogni Educativi Speciali della classe. Infine, la ricerca si è proposta di modellizzare i comportamenti osservati attraverso l'identificazione di cluster di interazione distinti, al fine di mappare le diverse posture professionali che i futuri docenti assumono nel dialogo riflessivo con le tecnologie generative, intese come catalizzatori per lo sviluppo di una rinnovata assessment literacy.

3. Disegno sperimentale: un approccio mixed-methods alla co-progettazione assistita

Lo studio adotta un disegno di ricerca a metodi misti di tipo esplicativo sequenziale (QUAN-qual), basato sull'analisi dei dati raccolti tramite un questionario CAWI (Computer Assisted Web Interviewing) somministrato a un campione di 191 studenti del Corso di Laurea in Scienze della Formazione Primaria, frequentanti il Laboratorio di Tecnologie Didattiche. La procedura di ricerca ha preso avvio all'interno del contesto laboratoriale, dove i partecipanti sono stati coinvolti in attività di co-progettazione di rubriche valutative supportate da sistemi di Intelligenza Artificiale. L'intero *workflow* di analisi è stato successivamente sviluppato nell'ambiente statistico RStudio, integrando variabili quantitative e narrazioni riflessive per monitorare l'esperienza in modo multidimensionale. La Tabella 1 illustra l'operazionalizzazione delle variabili e la selzione degli item del questionario analizzati nel presente contributo.

Dimensione analitica	**Item**	**Tipologia domanda**
Profilo professionale	Posizione lavorativa	Risposta chiusa
Percezione di utilità	Su una scala da 1 ("per nulla utile") a 10 ("estremamente utile"), quanto ritieni utile il dialogo con l'IA nella fase di progettazione delle attività?	Likert 10 punti
Negoziazione trasformativa	Su una scala da 1 ("per nulla") a 10 ("molto"), quanto hai modificato la tua idea iniziale grazie al confronto con l'IA?	Likert 10 punti
Teacher Agency	Su una scala da 1 ("solo il mio giudizio") a 10 ("ho seguito interamente l'IA"), quanto ti sei affidato/a al tuo giudizio rispetto alle proposte dell'IA?	Likert 10 punti
Assessment Literacy	Su una scala da 1 ("per nulla utile") a 10 ("molto utile"), come valuteresti il supporto ricevuto dall'IA nella strutturazione della valutazione?	Likert 10 punti
Intenzionalità inclusiva	Inserisci una breve contestualizzazione della classe scelta (numero alunni, presenza di BES, eventuali criticità o punti di forza, ecc.)	Risposta aperta
Riflessività pedagogica	Descrivi un esempio concreto in cui il confronto con l'IA ha modificato o arricchito uno degli elementi della tua progettazione. Quale ragionamento hai seguito?	Risposta aperta
Evoluzione della competenza	In che modo il confronto con l'IA ti ha aiutato a ripensare la valutazione del percorso didattico?	Risposta aperta

Tab. 1 – Operazionalizzazione delle variabili e struttura del questionario

Nella fase quantitativa, l'indagine ha impiegato scale Likert a dieci punti volte mappare i pattern statistici relativi alla relazione tra utilità percepita dell'IA nel design valutativo, l'effettivo grado di modifica delle idee progettuali originali e il livello di affidamento al proprio giudizio critico.

Nonostante l'omogeneità del target, composto interamente da docenti in formazione, è stato applicato il test di Welch per gestire le diverse dimensioni dei sottogruppi professionali emersi durante l'analisi, unitamente al coefficiente di Spearman per indagare il nesso tra l'agency docente, intesa come affidamento al proprio giudizio professionale, e la propensione alla revisione didattica.

L'identificazione dei profili professionali è stata realizzata attraverso una Cluster Analysis non gerarchica (algoritmo K-means) condotta nell'ambiente statistico RStudio. Come variabili di input (features) sono stati selezionati gli item quantitativi relativi all'utilità percepita, all'entità della modifica progettuale e al grado di agency. Il numero ottimale di cluster (k=3) è stato determinato attraverso la combinazione del metodo del gomito (Elbow Method) e del coefficiente di Silhouette, che hanno evidenziato una struttura robusta e coerente dei dati. L'applicazione di questa procedura ha permesso di segmentare i futuri docenti in profili di interazione tecnologica distinti (Sperimentatori, Mediatori, Prudenti), superando le semplici variabili anagrafiche a favore di una categorizzazione basata sulle posture esplorative adottate.

La successiva fase qualitativa, volta a spiegare e sostanziare i risultati numerici, ha utilizzato un'analisi tematica riflessiva (Braun & Clarke, 2006) sulle risposte aperte fornite dai corsisti al termine dell'esperienza laboratoriale, raccolte mediante il medesimo strumento CAWI. Si precisa che l'analisi si è focalizzata sulle rappresentazioni riflessive del processo di progettazione, consentendo di approfondire il ragionamento pedagogico e le strategie di negoziazione critica che gli studenti hanno attivato nel filtrare i suggerimenti algoritmici, ponendo in luce come la competenza valutativa in formazione venga mediata dall'uso dell'IA. L'integrazione delle evidenze ha infine permesso di far emergere la centralità dei temi dell'inclusione e della personalizzazione, che rappresentano i principali driver della riflessione critica documentata dai partecipanti nel dialogo con l'algoritmo.

4. Evidenze empiriche: pattern di interazione e scaffolding cognitivo

Il campione d'indagine è costituito da 191 studenti e riflette di genere tipica dei percorsi formativi per l'area pedagogica, evidenziando una prevalenza della componente femminile (97,4%, n=186) rispetto a quella maschile (2,6%, n=5). Sot-

to il profilo anagrafico, il gruppo manifesta un'elevata omogeneità, con il 60% dei soggetti collocato nella fascia *under* 24. Nonostante la natura *pre-service* del campione, con il 70,7% dei partecipanti non ancora stabilmente inserito nei ruoli docenti, emerge una significativa familiarità con l'Intelligenza Artificiale (93,7%), sebbene questa risulti confinata a un uso prevalentemente esplorativo e non ancora declinata in termini di competenza valutativa professionale. Tale evidenza configura un profilo di alfabetizzazione digitale generale che necessita di una mediazione pedagogica specifica per evolvere in reali competenze di progettazione assistita.

Le analisi descrittive rivelano una valutazione ampiamente positiva circa l'utilità del dialogo con l'IA nella fase di *design* (M = 8,30; SD = 1,28). Il supporto percepito, come illustrato nella Figura 1, si concentra in modo prevalente sulla strutturazione della valutazione (M = 7,87) e sulla pianificazione delle attività, corroborando l'ipotesi che l'algoritmo agisca come un efficace riduttore del carico cognitivo nei compiti caratterizzati da un'elevata complessità procedurale.

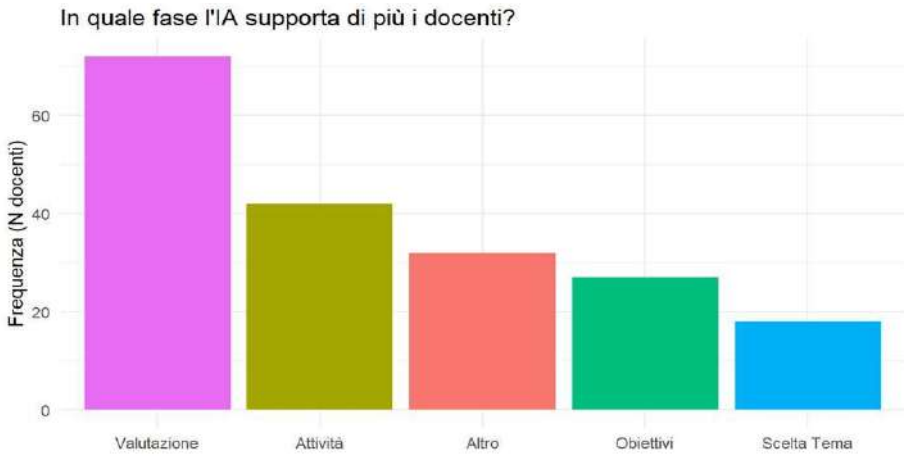


Fig. 1 – Supporto dell'IA per fase progettuale. Distribuzione delle frequenze relative all'utilità percepita nelle diverse fasi del design didattico

Tuttavia, si osserva uno scarto funzionale tra l'utilità dichiarata e l'effettivo grado di modifica del progetto originale (M = 6,14; SD = 2,23). Questa discrepanza suggerisce che l'IA non operi in termini di sostituzione dell'istanza pedagogica, ma piuttosto come uno strumento di *scaffolding* che i futuri docenti utilizzano per raffinare la propria visione iniziale senza cedere la regia del processo.

A sostegno di tale lettura, l'applicazione del test di Welch non ha rilevato discrepanze statisticamente significative ($p = 0,60$) tra soggetti con o senza pregresse esperienze di insegnamento, postulando che la tecnologia generativa sia percepita come una risorsa abilitante trasversale. Parallelamente, la correlazione di Spearman ha evidenziato un legame positivo significativo tra l'affidamento al proprio giudizio critico ($M = 6,80$) e la propensione alla negoziazione trasformativa con l'output algoritmico ($rho = 0,31$; $p < .001$). Sotto il profilo della progettazione inclusiva e della valutazione critica dello strumento, i dati sintetizzati nella Tabella 2 evidenziano come la maggioranza dei partecipanti (78,05%) abbia orientato la co-progettazione verso strategie per studenti con Bisogni Educativi Speciali. Emerge, parallelamente, una chiara consapevolezza dei benefici legati alla personalizzazione (n=109), controbilanciata dal timore di un possibile raffreddamento della relazione educativa (n=48).

Dimensione analizzata	**Nuclei tematici prevalenti**	**Occorrenze (n)**
Focus inclusione	Bisogni Educativi Speciali (BES)	86
	Gestione difficoltà di apprendimento	67
	Disturbi Specifici dell'Apprendimento (DSA)	47
	Strategie di personalizzazione e UDL	30
Vantaggi percepiti	Potenziamento dell'inclusività	109
	Ottimizzazione dei tempi di progettazione	74
	Supporto metodologico e guida procedurale	61
	Generazione di spunti creativi e idee	29
Rischi evidenziati	Raffreddamento della relazione educativa	48
	Diminuzione del senso critico e pigrizia cognitiva	35
	Presenza di bias e pregiudizi algoritmici	32
	Standardizzazione e omologazione del design	30

Tab. 2 – Analisi delle occorrenze tematiche: focus sull'inclusione, vantaggi percepiti e rischi rilevati nella co-progettazione assistita (N=191)

Attraverso l'applicazione dell'algoritmo di partizionamento *K-means*, è stata elaborata una tassonomia del comportamento tecnologico dei partecipanti, evidenziata nella Figura 2.

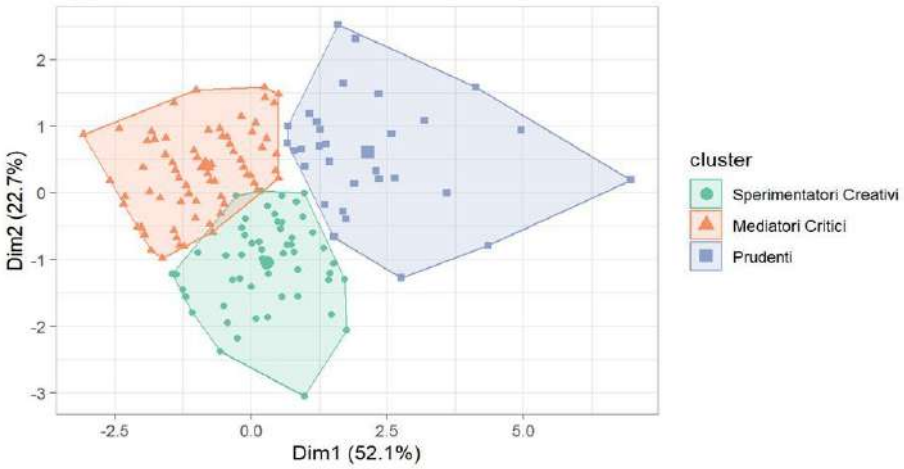


Fig. 2 – Profili di interazione docente-IA. Rappresentazione spaziale dei tre cluster identificati tramite algoritmo K-means: Sperimentatori Creativi, Mediatori Critici e Prudenti

La clusterizzazione ha permesso di identificare tre posture distinte: un primo gruppo di "Sperimentatori creativi" (34%), caratterizzati da un alto impatto trasformativo; una maggioranza di "Mediatori critici" (42%), che utilizza l'IA in modo strategico mantenendo un controllo rigoroso sul *design* pedagogico; e una quota di "Prudenti" (24%), contraddistinta da una bassa propensione alla revisione didattica. Il legame logico tra questa segmentazione e la valutazione dello strumento emerge chiaramente analizzando la Figura 3, che illustra il profilo qualitativo del feedback ricevuto. Il radar chart mostra infatti una copertura omogenea ed elevata in tutte le dimensioni (M T7,7), dalla pertinenza al valore formativo. Questa invarianza nella qualità dello stimolo tecnologico è un dato cruciale: essa dimostra che le differenze rilevate nei cluster non dipendono da una variabilità nelle performance dell'IA, bensì dalle diverse modalità di esercizio dell'*agency* docente da parte dei futuri docenti.

Fig. 3 – Radar chart basato sulla valutazione media (scala 1-10) delle dimensioni di pertinenza, chiarezza, adattabilità e valore formativo

5. Interpretare l'Agency docente nell'ecosistema valutativo post-digitale

L'analisi integrata dei risultati suggerisce che l'integrazione dell'Intelligenza Artificiale Generativa (GenAI) nella progettazione delle rubriche valutative non si configuri come una mera automazione tecnica, ma operi come un complesso dispositivo di *scaffolding* cognitivo. L'elevata utilità percepita dai futuri docenti (M = 8,30), contrapposta a una modifica solo moderata delle proposte originali (M = 6,14), indica che il dialogo con l'algoritmo non annulla l'intenzionalità del progettista, ma la corrobora attraverso un processo di confronto dialettico. In questo scenario, l'IA agisce propriamente come un partner riflessivo, uno specchio critico che permette al futuro docente di oggettivare le proprie scelte valutative e di raffinarle senza delegare la responsabilità decisionale alla macchina.

Il cuore interpretativo della ricerca risiede nella convergenza tra l'invarianza qualitativa dello stimolo tecnologico e la variabilità della risposta umana. Come dimostrato dal profilo multidimensionale del radar chart (Figura 3), l'IA ha fornito un feedback considerato uniformemente di alto livello in termini di pertinenza, chiarezza e valore formativo. Eppure, a fronte di questo input costante, la *cluster analysis* (Figura 2) ha rivelato una segmentazione profonda nelle posture professionali. Questo fenomeno è interpretabile attraverso il costrutto di agency docente: la diversità dei profili (Sperimentatori, Mediatori, Prudenti) non è imputabile a variazioni nelle performance dell'IA, ma riflette i diversi gradi di autonomia, visione pedagogica e resistenza critica dei futuri docenti. Il gruppo maggioritario dei "Mediatori Critici" (42%) incarna perfettamente l'ideale di un professionista riflessivo che, pur riconoscendo il valore aggiunto dell'algoritmo, ne filtra i suggerimenti attraverso il setaccio delle competenze pedagogiche acquisite nel percorso accademico.

Un ulteriore elemento di riflessione scaturisce dall'orientamento tematico documentato nella Tabella 2. La spiccata focalizzazione dei prompt sui temi dell'inclusione suggerisce che l'IA sia percepita come un mediatore privilegiato per operazionalizzare l'Universal Design for Learning (UDL). La tecnologia sembra dunque colmare lo scarto tra la teoria dell'inclusione e la sua difficile traduzione pratica nella progettazione di rubriche e attività personalizzate. Tuttavia, la consapevolezza dei rischi legati al possibile affievolimento della relazione educativa e alla standardizzazione della progettazione segnala che i futuri docenti mantengono una sana vigilanza etica.

6. Conclusione e traiettorie evolutive dell'identità professionale

Il presente studio ha permesso di analizzare criticamente l'intersezione tra la formazione iniziale dei docenti e le potenzialità offerte dall'Intelligenza Artificiale Generativa, evidenziando come quest'ultima possa configurarsi come un potente dispositivo di *scaffolding* pedagogico. Le evidenze raccolte dimostrano che, lungi dal produrre una deriva di automazione passiva, l'uso dell'IA all'interno di un contesto laboratoriale strutturato sollecita una negoziazione attiva e riflessiva. Il dato più rilevante risiede nel consolidamento della *teacher agency* (Priestley et al., 2015): i futuri docenti non si limitano a recepire l'output algoritmico, ma lo filtrano attraverso il setaccio della competenza valutativa e dell'orientamento inclusivo, con una spiccata attenzione alla personalizzazione per i Bisogni Educativi Speciali.

In termini di ulteriori sviluppi, la ricerca apre la strada a diverse linee di indagine. In primo luogo, appare prioritario strutturare studi longitudinali capaci di monitorare come le posture identificate (Sperimentatori, Mediatori, Prudenti) si evolvano nel passaggio dalla formazione accademica all'immissione in ruolo, al fine di verificare la tenuta della *agency* docente di fronte alle pressioni strutturali del contesto scolastico reale. Parallelamente, si rende necessaria una riflessione più ampia sulle implicazioni etiche e deontologiche dell'uso dell'IA nella valutazione, esplorando modalità di formazione che promuovano una *AI Literacy* ermeneutica, capace di prevenire rischi di omologazione del design didattico. Infine, i risultati suggeriscono l'opportunità di scalare il modello del Laboratorio di Tecnologie Didattiche ad altri contesti universitari, promuovendo la creazione di *repository* di prompt pedagogicamente validati che possano fungere da base comune per una progettazione valutativa sempre più equa, inclusiva e tecnologicamente consapevole.

Contributi degli autori

Il presente contributo è frutto di una collaborazione tra gli autori. Nello specifico, Viviana Vinci è autrice dei paragrafi 1, 5 e 6 mentre Pierangelo Berardi è autore dei paragrafi 2, 3 e 4. Entrambi gli autori hanno contribuito alla stesura dell'articolo e ne hanno revisionato e approvato la versione finale.

Riferimenti bibliografici

Boud, D., & Molloy, E. (Eds.) (2013). *Feedback in higher and professional education: Understanding it and doing it well*. Routledge.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Carless, D., & Winstone, N. (2023). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, *28*(1), 150–163. https://doi.org/10.1080/13562517.2020.1782372

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/-10.2307/249008

Fawns, T. (2022). An entangled pedagogy: Looking beyond the pedagogy—technology dichotomy. *Postdigital Science and Education*, *4*(3), 711–728. https://doi.org/-10.1007/s42438-022-00302-7

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Perla, L., Agrati, L. S., & Vinci, V. (2021). Valutare o valorizzare lo sviluppo professionale dell'insegnante: evidenze di ricerca per un dibattito. *Mizar. Costellazione di pensieri, 15*, 236-243.

Priestley, M., Biesta, G. J. J., & Robinson, S. (2015). *Teacher agency: An ecological approach*. Bloomsbury Academic.

Selwyn, N. (2019). *Should robots replace teachers? AI and the Future of Education*. (1st ed.) Polity Press.

Shulman, L. S. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review*, 57, 1-22. http://dx.doi.org/10.17763/haer.57.-1.j463w79r56455411

Stiggins, R. J. (1991). Assessment Literacy. *The Phi Delta Kappan*, *72*(7), 534–539.

Striano, M., Melacarne, C., & Oliverio, S. (2018). *La riflessività in educazione. Prospettive, modelli, pratiche*. Brescia: Scholè.

Thanh, B. N., Vo, D. T. H., Nhat, M. N., Pham, T. T. T., Trung, H. T., & Xuan, S. H. (2023). Race with the machines: Assessing the capability of generative AI in solving authentic assessments. Australasian Journal of Educational Technology, 39(5), 59-81. https://doi.org/10.14742/ajet.8902

Vinci, V., & Berardi, P. (2025). Between enthusiasm and resistance: perceptions of ai in teacher education. *Giornale italiano di educazione alla salute, sport e didattica inclusiva, 9*(2). https://doi.org/10.32043/gsd.v9i2.1430

Williams, A. (2025). Integrating artificial intelligence into higher education assessment. *Intersection: A Journal at the Intersection of Assessment and Learning, 6*(1), 128-154. https://doi.org/10.61669/001c.131915

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100. https://doi.org/-10.1111/j.1469-7610.1976.tb00381.x

II.22
L'IA : un mediatore digitale nel processo di feedback per la formazione di insegnanti pre-service e a scuola

Antonella Montone
Università degli studi di Bari Aldo Moro
antonella.montone@uniba.it
Rosalba Romito
Università degli studi di Bari Aldo Moro
Michele G. Fiorentino
Università degli studi di Bari Aldo Moro
Michele Romita
Università degli studi di Bari Aldo Moro

Abstract

Il contributo indaga il ruolo dell'Intelligenza Artificiale come mediatore digitale nel processo di feedback in matematica, nella formazione iniziale e a scuola. Lo studio analizza una sperimentazione in tre fasi (risoluzione di un problema aperto, interazione con ChatGPT tramite prompt strutturato, discussione collettiva), interpretata alla luce della Mediazione Semiotica e della valutazione formativa. L'analisi di due casi evidenzia che l'IA sostiene l'esplicitazione argomentativa e la riflessione metacognitiva, ma mostra limiti nella diagnosi dell'errore concettuale, richiedendo una mediazione docente.

Parole chiave: AI; teoria della mediazione semiotica; formazione insegnanti; feedback; valutazione formativa.

This paper investigates Artificial Intelligence as a digital mediator in formative feedback processes in mathematics, both in teacher education and school contexts. A three-phase design (open problem solving, structured interaction with ChatGPT, collective discussion) was implemented and interpreted through Semiotic Mediation Theory and formative assessment. The analysis of two cases shows that AI enhances argumentative explicitation and metacognitive reflection, but reveals limits in diagnosing conceptual errors, thus requiring explicit teacher mediation.

Keywords: AI; theory of semiotic mediation; teacher education; feedback; formative assessment.

1. Introduzione

Negli ultimi anni, il tema del feedback ha assunto un ruolo centrale nel dibattito pedagogico e didattico, in particolare nell'ambito dell'educazione matematica. Numerosi studi evidenziano come la qualità e la tempestività del feedback influenzino in modo significativo i processi di apprendimento, favorendo la consapevolezza metacognitiva degli studenti e sostenendo lo sviluppo di competenze argomentative e di problem solving (Hattie & Timperley, 2007; Black & Wiliam, 2009). Tuttavia, nella pratica scolastica quotidiana, l'implementazione di un feedback efficace risulta spesso problematica a causa di vincoli organizzativi, del numero elevato di studenti e della difficoltà di gestire in tempi brevi la restituzione e la discussione delle produzioni. In questo scenario si inserisce il crescente interesse verso l'Intelligenza Artificiale (IA) generativa, che negli ultimi anni ha mostrato potenzialità rilevanti in ambito educativo (Celik et al., 2022; Ji et al., 2023; Salas-Pilco et al., 2023). In particolare, sistemi basati su modelli linguistici di grandi dimensioni, come ChatGPT, consentono interazioni dialogiche complesse, capaci di adattarsi alle risposte dell'utente e di fornire suggerimenti articolati. Tali caratteristiche aprono nuove prospettive per l'uso dell'IA non solo come strumento di supporto operativo, ma come possibile risorsa didattica integrata nei processi di insegnamento e apprendimento. Il presente contributo si colloca in questo quadro e propone di interpretare l'IA come mediatore digitale all'interno del processo di feedback in Matematica, assumendo come riferimento teorico la Teoria della Mediazione Semiotica (Bartolini Bussi & Mariotti, 2008) e le pratiche di valutazione formativa (Black & Wiliam, 2009). L'obiettivo è analizzare le interazioni dello studente con l'IA, integrata nella risoluzione di un problema aperto e nella successiva rielaborazione delle soluzioni, stabilire quanto tali interazioni possano incidere sui processi argomentativi dello studente stesso e supportare il lavoro dell'insegnante nella gestione della discussione matematica (Bartolini Bussi, 1989).

2. Quadro teorico e integrazione dell'IA

La progettazione e l'analisi dell'attività oggetto di questo studio si fondano su un quadro teorico che intreccia la Teoria della Mediazione Semiotica e il modello della valutazione formativa. Nella Teoria della Mediazione Semiotica l'apprendimento è inteso come un processo socialmente e culturalmente mediato, nel quale la costruzione dei significati matematici avviene attraverso l'interazione tra soggetti, linguaggi e artefatti. In questa prospettiva, l'artefatto non è riducibile

a un semplice strumento operativo, ma assume lo status di oggetto culturale, avente uno specifico potenziale semiotico che orienta l'attività degli studenti e favorisce la produzione di segni personali. Tali segni, inizialmente legati all'azione e all'esperienza concreta con l'artefatto, possono progressivamente evolvere verso significati matematici più astratti e condivisi, fino a essere istituzionalizzati nel discorso matematico scolastico. Il processo di insegnamento-apprendimento viene così inteso come un percorso di trasformazione semiotica, che non avviene spontaneamente, ma richiede una progettazione intenzionale.

All'interno di questo quadro, il ruolo dell'insegnante risulta centrale: egli agisce come mediatore e moderatore dei processi di costruzione dei significati matematici in gioco, selezionando gli artefatti, progettando i compiti e orchestrando le interazioni, affinché i segni prodotti dagli studenti possano evolversi e da segni personali svilupparsi in segni matematici.

In questa prospettiva, la valutazione formativa assume un ruolo fondamentale. Essa si configura come una pratica integrata, finalizzata a sostenere e orientare l'apprendimento, piuttosto che a certificare esclusivamente i risultati in un momento separato o successivo all'attività didattica (Black & Wiliam, 2009; Castoldi, 2012). La valutazione formativa si fonda sulla raccolta sistematica di informazioni sui processi di apprendimento degli studenti e sull'uso di tali informazioni per guidare le decisioni didattiche dell'insegnante e favorire l'autoregolazione degli studenti stessi. In questo contesto, il feedback rappresenta uno snodo cruciale: un feedback efficace contribuisce a rendere visibili le strategie adottate, le difficoltà concettuali e le possibili direzioni di miglioramento e non esclusivamente a segnalare l'errore o la correttezza di una risposta (Hattie & Timperley, 2007). In matematica, tale funzione è particolarmente delicata, poiché il feedback deve tenere insieme dimensioni concettuali, procedurali e argomentative, evitando una riduzione dell'apprendimento a mera esecuzione tecnica.

La Teoria della Mediazione Semiotica interpreta la valutazione formativa come un ulteriore dispositivo di mediazione, attraverso il quale l'insegnante attribuisce senso ai segni prodotti dagli studenti e li accompagna nella rielaborazione del proprio pensiero. Valutazione e mediazione concorrono così alla costruzione del significato matematico, configurandosi come pratiche profondamente intrecciate. In questo spazio teorico integriamo l'Intelligenza Artificiale, intesa come artefatto didattico. L'IA presenta caratteristiche specifiche che arricchiscono quelle di un artefatto così come inteso nell'ambito della TMS. Infatti, l'IA è in grado di interagire linguisticamente con lo studente, di porre domande, di suggerire strategie alternative e di sollecitare chiarimenti, configurandosi come un mediatore semiotico dinamico.

In questa prospettiva, l'IA si pone come uno strumento che può supportare il processo di feedback e favorire la rielaborazione delle soluzioni e la riflessione metacognitiva degli studenti, e non va intesa come un sostituto dell'insegnante né come una fonte di verità matematica. Il suo contributo risulta particolarmente significativo nella fase di restituzione, in cui il feedback è più tempestivo e sostenibile, senza snaturarne la funzione formativa. Tuttavia, affinché l'IA svolga effettivamente una funzione di mediazione e non si riduca a un correttore automatico, è necessario che il suo utilizzo sia inserito in una progettazione didattica consapevole e guidata. L'insegnante mantiene un ruolo centrale nella definizione dei compiti, nella selezione delle modalità di interazione con l'IA e nella gestione della discussione collettiva, affinché le interazioni con il mediatore semiotico dinamico possano essere ricondotte a significati matematici condivisi. Alla luce di queste considerazioni, il presente contributo si pone l'obiettivo di analizzare in che modo l'interazione con l'IA possa influire sulla rielaborazione delle soluzioni di un problema matematico aperto, e in che misura possa supportare l'insegnante nella gestione del feedback, al fine di caratterizzare il ruolo dell'Intelligenza Artificiale come mediatore semiotico dinamico.

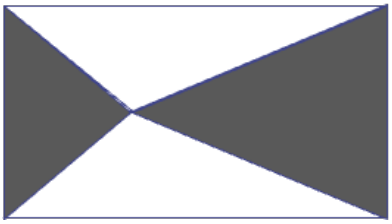
3. Metodologia

3.1 Scelta del problema aperto

La sperimentazione è stata condotta in contesti scolastici di scuola primaria e secondaria, nell'ambito di attività curricolari di matematica, e in un corso di formazione per insegnanti pre-service di Scienze della Formazione Primaria, nell'ambito del corso di didattica della matematica. Gli studenti hanno lavorato in coppie, al fine di favorire il confronto e la negoziazione delle strategie risolutive.

Il compito proposto è un problema aperto di geometria piana, noto come *problema dell'eredità*:

> *"Due fratelli ereditano un terreno di forma rettangolare. Per dividerlo in due parti della medesima estensione, un conoscente suggerisce loro di piantare un palo in un punto qualsiasi del terreno e congiungerlo ai quattro paletti infissi nei quattro vertici del terreno rettangolare.*
> *Uno dei fratelli prenderà la parte colorata in grigio nel disegno, l'altro la rimanente.*
> *Le due parti sono davvero uguali?*
> *Giustificate il vostro ragionamento"*



Il problema è stato selezionato perché ammette differenti strategie risolutive, in modo tale da offrire la possibilità agli studenti di individuare almeno una soluzione. I problemi aperti rivestono un ruolo fondamentale nella didattica della matematica in quanto stimolano processi di pensiero non riducibili a una singola procedura o a un'unica soluzione. Diversamente dai problemi chiusi, che tendono a privilegiare l'applicazione di algoritmi noti, i problemi aperti richiedono agli studenti di ragionare, riflettere e giustificare più strategie possibili, favorendo lo sviluppo di competenze come la creatività, il pensiero critico e la flessibilità cognitiva. Diversamente dai problemi chiusi, che tendono a privilegiare l'applicazione di algoritmi noti, i problemi aperti richiedono agli studenti di ragionare, riflettere e giustificare più strategie possibili, favorendo lo sviluppo di competenze come la creatività, il pensiero critico e la flessibilità cognitiva (Schoenfeld, 1985). Queste tipologie di compiti consentono inoltre agli insegnanti di ottenere maggiori informazioni sulle strategie risolutive adottate dagli studenti, sulle loro idee e sulle misconcezioni, offrendo così evidenze formative utili per orientare l'insegnamento. Tali caratteristiche rendono il problema aperto uno strumento didattico utile per valorizzare l'esperienza matematica come un processo dinamico di costruzione di significato, in cui gli studenti sono chiamati a formulare ipotesi, valutare strategie alternative e giustificare le proprie scelte, azioni caratterizzanti i processi di valutazione formativa.

3.2 Struttura della sperimentazione e raccolta dati

La sperimentazione si è articolata in tre fasi:

– Fase 1 – Risoluzione del problema: gli studenti hanno risolto il problema lavorando in coppie, producendo una prima soluzione argomentata. Le risposte sono state raccolte tramite Google Moduli, al fine di documentare le strategie adottate e le argomentazioni sviluppate.

– Fase 2 – Interazione con l'IA: In questa fase gli studenti hanno potuto interagire con ChatGPT utilizzando un prompt comune, che invitava gli studenti a rivedere e migliorare la propria soluzione iniziale sulla base dei feedback ricevuti.

Il prompt strutturato per gli studenti di scienze della formazione primaria era:

> *"Sono un* student* di Scienze della Formazione Primaria, e ho risolto il seguente problema di matematica. Vorrei che mi aiutassi a riflettere sui passaggi della mia risoluzione senza fornirmi delle risposte, ma dandomi dei feedback formativi. La traccia del problema è "Due fratelli ereditano un terreno di forma rettangolare. Per dividerlo in due parti della medesima estensione, un conoscente suggerisce loro di piantare un palo in un punto qualsiasi del terreno e congiungerlo ai quattro paletti infissi nei quattro vertici del terreno rettangolare.*
> *Uno dei fratelli prenderà la parte colorata in grigio nel disegno, l'altro la rimanente.*
> *Le due parti sono davvero uguali?*
> *Giustificate il vostro ragionamento".*
> *La mia risoluzione è: (inserire la propria risoluzione con l'immagine)"*

Gli studenti hanno prodotto una nuova versione della soluzione e hanno allegato la trascrizione dell'interazione con l'IA, indicando se le modifiche apportate fossero assenti, parziali o sostanziali. Per gli studenti di scuola primaria e secondaria, il prompt è analogo, e si modifica nella descrizione iniziale "*sono uno studente di scuola primaria/secondaria e ho risolto il seguente problema di matematica…*"

– Fase 3 – Discussione collettiva: le soluzioni sono state oggetto di una discussione collettiva guidata dall'insegnante. Le produzioni e le interazioni con l'IA hanno costituito materiale di lavoro per il rilancio delle idee e per l'istituzionalizzazione dei significati matematici.

Le scelte di design della sperimentazione sono state orientate a mantenere costante la variabile-compito, proponendo il medesimo problema aperto in contesti differenti per rendere confrontabili i processi di rielaborazione e le interazioni con l'IA. L'ipotesi di lavoro era che un'interazione dialogica con feedback non risolutivo potesse favorire l'esplicitazione argomentativa e la revisione delle strategie, con effetti differenziati in base alla correttezza iniziale delle soluzioni.

La fornitura di un prompt comune ha avuto funzione di standardizzazione dell'interazione, riducendo la variabilità delle richieste rivolte al sistema. Il prompt è stato progettato per orientare l'IA verso feedback formativo, riflessivo e non direttivo, coerentemente con il quadro teorico della mediazione semiotica e della valutazione formativa.

4. Analisi dell'interazione con IA: il caso di una soluzione non corretta e il caso di una soluzione corretta

Si riporta qui di seguito l'analisi di due casi nell'ambito del corso di scienze della formazione primaria.

4.1 Analisi di caso di una soluzione non corretta

Il primo protocollo analizzato è stato prodotto da una coppia di studentesse che ha inizialmente fornito una risposta non corretta al problema dell'eredità. L'interesse per tale caso è legato al fatto che le due studentesse non sono riuscite ad individuare l'invarianza dell'area in modo corretto già nella prima fase, così come è accaduto ad altri studenti. Questo rende il caso particolarmente significativo dal punto di vista interpretativo, perché mostra in che modo gli studenti che individuano una soluzione non corretta possano interagire con l'IA e quali criticità emergano nel processo.

La risposta iniziale delle due studentesse consisteva nell'affermazione:

> "*Le due parti non sono uguali perché non sappiamo se il punto in cui viene messo il palo sia effettivamente al centro del terreno*."

L'argomentazione si fonda sull'idea che l'uguaglianza delle due superfici dipenda necessariamente dalla posizione centrale del punto interno; si tratta di un errore concettuale, riconducibile a un fraintendimento della struttura geometrica della partizione: non sono i singoli triangoli a dover essere uguali, ma le coppie di triangoli opposti, la cui area combinata è costante indipendentemente dalla posizione del punto.

L'analisi del protocollo mostra che il feedback prodotto da ChatGPT, pur non individuando l'errore concettuale contenuto nella risposta iniziale, presenta alcune caratteristiche che si avvicinano alla definizione teorica di feedback formativo. L'IA, infatti:

– non fornisce la soluzione, ma invita a rivedere e rielaborare la motivazione;
– struttura il feedback in sezioni (cosa funziona / cosa migliorare);
– suggerisce strategie di argomentazione alternative (uso del disegno, confronto tra aree, controesempi);
– formula domande guida che sollecitano riflessione e ristrutturazione cognitiva.

Questi elementi sono coerenti con l'idea di feedback, inteso come stimolo alla rielaborazione autonoma, e riflettono l'impostazione della valutazione formativa intesa come accompagnamento al ragionamento e alla rielaborazione consapevole delle soluzioni.

L'aspetto più rilevante del protocollo riguarda il comportamento dell'IA nella prima interazione. Nonostante la risposta fosse errata, l'IA non ha riconosciuto l'errore concettuale e ha invece fornito un feedback di incoraggiamento, confermando alcuni elementi della risposta:

> "*Cosa funziona nella tua risposta*: *Hai riconosciuto il punto chiave: la posizione del palo è rilevante… La tua osservazione è prudente: non affermi qualcosa senza motivarla (dici "non sappiamo se…"), il che è corretto in un contesto dimostrativo.*"

L'IA offre alle studentesse la possibilità di riflettere ulteriormente sulla risposta data e mette in rilievo un elemento fondamentale del problema: l'invarianza dell'equiestensione correlata alla posizione del palo. Inoltre, incoraggia le studentesse a procedere nella dimostrazione in un clima di supporto dialogico. Tuttavia, non è sempre in grado di discriminare tra risposta valida e risposta scorretta, e tende talvolta ad assumere una posizione confermativa.

In generale, l'intervento dell'IA mostra alcune criticità:

– il modello non ha riconosciuto l'errore di fondo presente nella risposta iniziale ("le due parti non sono uguali"), limitandosi a richiamare parti della soluzione, anziché problematizzarle;
– il feedback si concentra prevalentemente su aspetti strutturali e comunicativi ("la frase è troppo sintetica"), senza intervenire sul significato matematico;
– non viene fornito un confronto diretto tra la tesi esposta e la realtà matematica del problema, come avviene nella valutazione formativa docente-studente;
– manca una direzionalità precisa verso l'obiettivo di apprendimento, poiché l'IA, evitando di giudicare la correttezza, non chiarisce se la posizione scelta per il punto sia promettente o fuorviante.

Tuttavia, il protocollo presenta anche un esito significativo: attraverso la sequenza di domande e suggerimenti, le due studentesse, nella versione rielaborata, formulano una risposta corretta e argomentata. La nuova formulazione mostra una comprensione più profonda della struttura del problema:

> *"Le due parti non sono necessariamente uguali: lo sono se si prendono due triangoli con basi sui lati opposti (es triangoli sopra/sotto) perché l'altezza di uno e quella dell'altro sommate danno sempre l'altezza totale; quindi, la loro area insieme è sempre metà del rettangolo, indipendentemente da dove si trova il punto. Invece, non lo sono se si prendono due triangoli adiacenti (per esempio quelli a sinistra), allora le aree non sono in generale uguali, tranne nel caso in cui il palo sia esattamente al centro del rettangolo, cioè sul punto d'intersezione delle diagonali."*

Attraverso la frase "*quindi, la loro area insieme è sempre metà del rettangolo, indipendentemente da dove si trova il punto"* dimostrano di aver compreso la richiesta del problema e l'invarianza sottostante, espressa attraverso la parola "*indipendentemente*".

Nel commento finale, le studentesse attribuiscono esplicitamente la nuova proposta di soluzione ad alcune domande dell'IA. Infatti, ChatGPT fornisce loro il seguente suggerimento:

> "*Cosa migliorare / sviluppare:*
> *La frase attuale è troppo sintetica: affermare che «le due parti non sono uguali perché non sappiamo se il punto è il centro» non è ancora una giustificazione matematica. Serve spiegare perché la posizione del punto centrale (o non centrale) influisce sulle aree.*
> *Potresti trasformare l'argomento da una semplice affermazione a un ragionamento articolato: enuncia un'ipotesi chiara, mostra come ciò conduce a una differenza nelle aree, o costruisci un controesempio se possibile*."

Le due studentesse, in riferimento a tale suggerimento, replicano la loro soluzione con l'affermazione seguente:

> "*Le domande che si riferiscono al metodo analitico semplice, come guida per il ragionamento... Abbiamo poi argomentato il ragionamento di base*."

Questo elemento rafforza l'idea dell'IA come mediatore dialogico: non fornisce la soluzione, ma stimola riflessione e revisione dei ragionamenti.

In questo senso, l'interazione rivela una doppia natura dell'IA come media-

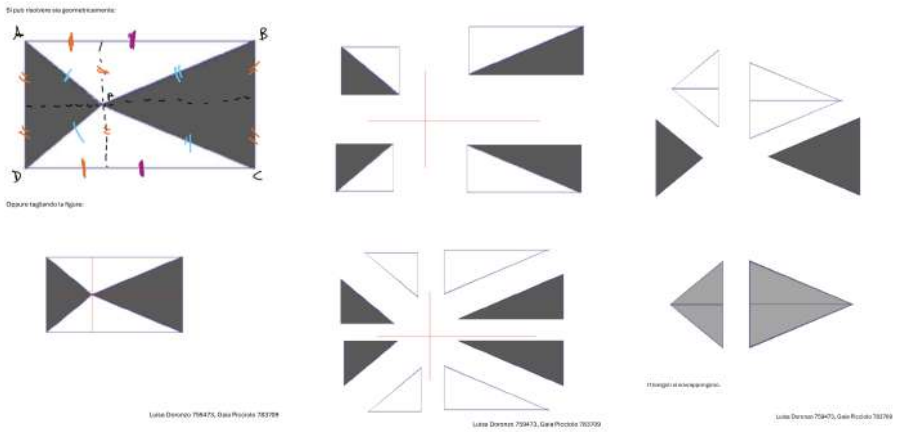
tore digitale. Sul piano dialogico, ChatGPT svolge effettivamente una funzione di scaffolding: invita alla rielaborazione, espande il campo di pensiero delle studentesse, introduce strumenti concettuali nuovi e li fa emergere senza rivelare la soluzione. Su questo piano, il comportamento dell'IA è pienamente allineato con la prospettiva formativa.

Sul piano epistemico, però, l'IA non svolge il ruolo di garante disciplinare: non individua errori, non corregge distorsioni concettuali e non aiuta a distinguere ciò che è matematicamente vero da ciò che è solo plausibile.

4.2 Analisi di una risposta corretta

Il secondo protocollo analizzato mostra un caso in cui gli studenti hanno fornito fin dall'inizio una soluzione corretta al problema dell'eredità. La risposta iniziale evidenzia una comprensione solida della proprietà geometrica sottostante: il rettangolo viene suddiviso in quattro triangoli collegando un punto interno ai vertici, e l'uguaglianza delle due regioni è sostenuta dal fatto che i triangoli opposti hanno basi su lati paralleli e altezze comuni. Le immagini prodotte dagli studenti confermano questa interpretazione: nella figura, la parte grigia è ricomposta in triangoli che si sovrappongono perfettamente a quelli della parte bianca, mostrando visivamente la coincidenza delle aree. A differenza del primo caso analizzato, qui gli studenti adoperano due strategie argomentative distinte: una di natura empirica e visiva (taglio e sovrapposizione dei triangoli) e una di natura geometrica deduttiva basata su parallelismo e altezze uguali. Questa doppia argomentazione mostra un livello di padronanza concettuale elevato e coerente con i criteri di correttezza della soluzione.

> "*Il problema può essere risolto sia geometricamente con una dimostrazione o tagliando la figura e sovrapponendo le varie parti.*"



Il feedback dell'IA, in questo caso, si configura come una forma di consolidamento formativo: il modello riconosce correttamente che il ragionamento svolto è fondato, ma invita gli studenti a esplicitare i principi geometrici che giustificano la sovrapposizione delle aree: il feedback sottolinea ad esempio la necessità di dichiarare che i triangoli opposti hanno stessa base e altezza perché hanno le basi su lati paralleli. Nella sezione dedicata alla spiegazione di cosa hanno modificato in seguito all'interazione con ChatGPT, gli studenti hanno scritto:

> "*Abbiamo inserito una parte discorsiva (con l'aggiunta di proprietà geometriche) che spiegava come avevamo ragionato...*"

L'interazione con l'IA ha prodotto un effetto significativo sulla versione riformulata della risposta. Gli studenti hanno trasformato una soluzione prevalentemente visiva in una più discorsiva e dimostrativa, esplicitando concetti prima impliciti, come il parallelismo dei lati e l'indipendenza della soluzione dalla posizione del punto.

L'AI gli suggerisce quanto segue:

> *"La tua risoluzione è corretta e mostra una buona padronanza concettuale.*
> *I punti da rafforzare riguardano esplicitare:*
> *le proprietà geometriche che usi;*
> *perché la sovrapposizione funziona sempre;*
> *perché il punto scelto può essere qualunque.*
> *Hai quindi un'ottima base: basta rendere un po' più "visibile" il ragionamento che hai in testa."*

Nella versione modificata scrivono infatti:

> "*Nel rettangolo i lati opposti sono paralleli: questo fa sì che i triangoli che hanno come base due lati opposti […] abbiano la stessa altezza […] sempre, qualunque sia il punto in cui si pianta il palo. Di conseguenza hanno la stessa area sempre.*"

È interessante notare che, la parola "*sempre*" utilizzata dall'IA nel suggerimento "*esplicitare perché la sovrapposizione funziona sempre*", induce gli studenti ad esplicitare il processo dimostrativo avendo ora ben chiaro il fatto che il problema chiede proprio di verificare che l'area della due parti sia equivalente sempre, qualunque sia il punto in cui è piantato il palo. Quindi diventano consapevoli della generalizzazione sottostante la richiesta del problema. Gli studenti elaborano una dimostrazione geometrica più dettagliata e logicamente corretta che, rispetto a quella iniziale, si arricchisce di nuovi elementi di matematizzazione e di passaggi logici ora espliciti.

Inoltre, mentre nel primo protocollo l'IA non è stata in grado di correggere un errore concettuale, qui si osserva la situazione opposta: l'IA opera con efficacia come facilitatore discorsivo e come mediatore linguistico, aumentando la trasparenza del processo matematico, pur senza aggiungere nuovi contenuti disciplinari. In sintesi, il primo protocollo evidenzia un possibile limite dell'IA nella diagnosi dell'errore concettuale, mentre il secondo suggerisce che, in presenza di soluzioni corrette, l'interazione possa sostenere processi di esplicitazione e raffinamento argomentativo. Pur non essendo generalizzabili, questi risultati indicano implicazioni didattiche rilevanti: l'IA sembra più affidabile come facilitatore discorsivo e metacognitivo che come strumento diagnostico, e richiede quindi una mediazione docente nella validazione epistemica dei contenuti.

5. Conclusioni

Il contributo ha mostrato come il confronto diretto tra valutazione umana e valutazione automatizzata possa costituire un potente dispositivo formativo per i docenti in formazione iniziale. L'Intelligenza Artificiale, lungi dall'essere assunta come soluzione tecnica o strumento sostitutivo del giudizio professionale, si configura piuttosto come un elemento perturbante capace di rendere visibili dimensioni implicite dell'agire valutativo. Le discrepanze emerse tra feedback umano e feedback generato dall'IA hanno sollecitato nei docenti processi di

riflessione critica sui criteri, sulle priorità pedagogiche e sulla responsabilità etica connessa alla valutazione.

I risultati permettono di riconoscere all'IA un potenziale supporto per gli aspetti formali e strutturali del feedback, ma ne individuano con chiarezza i limiti nel cogliere la complessità dell'apprendimento, il percorso dello studente e la dimensione relazionale che caratterizza l'atto valutativo. In questo senso, l'esperienza conferma la necessità di superare visioni dicotomiche – entusiasmo acritico o rifiuto difensivo – per promuovere modelli valutativi ibridi, nei quali l'IA sia integrata in modo intenzionale, trasparente e pedagogicamente orientato.

Dal punto di vista formativo, emerge con forza l'urgenza di includere percorsi strutturati di AI literacy critica nella formazione iniziale e continua degli insegnanti. Tali percorsi dovrebbero mirare non solo allo sviluppo di competenze operative, ma soprattutto alla costruzione di una consapevolezza epistemologica ed etica rispetto all'uso dell'IA nei processi educativi. In questa prospettiva, la valutazione resta un atto eminentemente umano, fondato sull'interpretazione, sulla responsabilità e sulla relazione educativa, che nessun sistema automatizzato può sostituire.

Infine, il contributo apre a sviluppi futuri di ricerca, orientati ad ampliare il campione, a esplorare differenze tra ambiti disciplinari e a indagare in modo più approfondito l'impatto dell'IA sull'identità professionale docente. In un contesto educativo attraversato da rapide trasformazioni tecnologiche, interrogare criticamente il ruolo dell'Intelligenza Artificiale nella valutazione non rappresenta una scelta opzionale, ma una responsabilità pedagogica.

Riferimenti bibliografici

Bartolini Bussi, M.G. (1989). La discussione collettiva nell'apprendimento della matematica. *Parte II, L'Insegnamento della Matematica e delle Scienze Integrate*, 12(5), 615–654.

Bartolini Bussi, M.G., & Mariotti, M.A. (2008). *Semiotic mediation in mathematics education*.

Black, P. & William, D. (2009). Developing the Theory of Formative Assessment. *Educational Assessment, Evaluation and Accountability*, 21(1)

Castoldi, M. (2012). *Valutare a scuola. Dagli apprendimenti alla valutazione di sistema* (pp. 1-364). Roma: Carocci.

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The Promises and Challenges of Artificial Intelligence for Teachers: a Systematic Review of Research. *TechTrends*, 66, 616-630. DOI: 10.1007/s11528-022-00715-y.

Salas-Pilco, S. Z., Xiao, K., & Hu, X. (2023). Correction: Salas-Pilco et al. Artificial Intelligence and Learning Analytics in Teacher Education: A Systematic Review. *Education Science*, 13(9), 897.
Schoenfeld, A. H. (1985). *Mathematical Problem Solving*. Academic Press.
Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1)

II PARTE

Area tematica 4. IA e feedback

II.23
L'AI generativa come supporto alla feedback literacy degli studenti[1]

Manuela Fabbri
Professoressa Associata in Didattica Generale e Pedagogia Speciale, Università degli Studi di Bologna, m.fabbri@unibo.it
Chiara Laici
Professoressa Associata in Didattica Generale e Pedagogia Speciale, Università degli Studi di Macerata, chiara.laici@unimc.it
Maila Pentucci
Professoressa Associata in Didattica Generale e Pedagogia Speciale, Università "d'Annunzio" di Chieti-Pescara
maila.pentucci@unich.it

Abstract

Lo studio esamina il potenziale dei sistemi generativi di IA nel processo di feedback universitario, con particolare attenzione al loro contributo allo sviluppo della feedback literacy. La ricerca, condotta con 62 studenti di area pedagogica, ha previsto la progettazione collaborativa di un compito, la sua generazione tramite un modello linguistico, il confronto tra i due prodotti e la ricezione di feedback automatizzato, seguita da momenti riflessivi individuali e di gruppo. I risultati mostrano che la comparazione con l'elaborato generato dall'IA è stata percepita dalla maggior parte degli studenti come un'opportunità di apprendimento metacognitivo. Il feedback automatico è stato valutato come rapido, chiaro e utile all'integrazione dei lavori, pur evidenziando limiti legati alla genericità e alla scarsa personalizzazione. Nel complesso, l'IA emerge come risorsa complementare alla mediazione docente, in grado di sostenere pratiche di feedback più sostenibili e favorire l'autoregolazione. L'esperienza suggerisce la necessità di sviluppare una cultura educativa dell'IA che integri efficacia algoritmica e sensibilità pedagogica.

1 Il capitolo si inserisce all'interno di un lavoro di ricerca sviluppato in forma collaborativa dalle autrici. I contributi sono attribuiti come segue: 1. *Introduzione e background* e *3. Risultati e discussione*: Chiara Laici; *2. Descrizione del percorso e metodologia di analisi* e *3.1 Comparative feedback*: Manuela Fabbri; *3.2 Il feedback trasformativo* e *3.3 Il feedback riflessivo*: Maila Pentucci. Il paragrafo *4. Conclusioni* è stato elaborato congiuntamente.

Parole chiave: Feedback Literacy; Feedback Loop; Feedback trasformativo, Large Language Models, Higher Education.

The study examines the potential of generative AI systems in the university feedback process, with a focus on their contribution to the development of feedback literacy. The research, conducted with 62 students in the field of education, involved the collaborative design of a task, its generation using a language model, a comparison between the two products, and the receipt of automated feedback, followed by individual and group reflection sessions. The results show that most students perceived the comparison with the AI-generated work as an opportunity for metacognitive learning. The automated feedback was evaluated as quick, clear, and useful for integrating the work, while highlighting limitations related to its generality and lack of personalization. Overall, AI emerges as a complementary resource to teacher mediation, capable of supporting more sustainable feedback practices and promoting self-regulation. The experience highlights the importance of developing an educational culture of AI that balances algorithmic efficiency with pedagogical sensitivity.

Keywords: Feedback Literacy; Feedback Loop; Trasformative Feedbck, Large Language Models, Higher Education.

1. Introduzione e background

Le potenzialità dell'uso dell'AI generativa (AIG) nei contesti didattici, in particolare nell'higher education, sono state indagate negli ultimi anni in maniera ampia e diversificata (Hwang & Chang, 2021; Morris et al., 2021), contribuendo a consolidare la consapevolezza che gli agenti artificiali non possano essere esclusi dagli ecosistemi formativi, ma debbano essere considerati elementi proattivi dell'infosfera (Floridi, 2025), con i quali gli esseri umani interagiscono e si confrontano quotidianamente.

In tale scenario, la natura conversazionale dei Large Language Models (LLM) ha favorito un crescente interesse verso il loro impiego nei processi di feedback. La letteratura ha evidenziato come questi sistemi possano essere un supporto per il docente nell'attivazione del *feedback loop*, garantendo tempestività nelle risposte (*immediate feedback*), scalabilità nei contesti didattici ad alta numerosità (*large classroom*) e abbiano un potenziale significativo nello sviluppo di pratiche di autovalutazione e *inner feedback* da parte degli studenti (Nicol, 2019; Carless, 2019).

Ciò consente oggi di avviare riflessioni di più ampia portata, che vadano oltre la descrizione dell'efficacia operativa dei LLM come strumento di feedback nella

pratica e si proiettino verso una ricognizione delle loro sfaccettature di agency e la loro portata interoggettiva (Latour, 1994). In particolare, è interessante esplorare la questione se l'AIG possa, non solo svolgere procedure, ma anche agire come facilitatore nella costruzione di competenze riflessive e come dispositivo di mediazione per l'incorporazione di posture negli studenti, attraverso affordances (es. urgenza di risposta, dialogicità, aspettativa di revisione) incoraggiate dall'agentività della macchina (Deleuze & Guattari, 1980).

Nel presente contributo si intende comprendere se e come, in un processo didattico strutturato su cicli di feedback ricorsivi e diversificati, l'AIG possa contribuire al consolidamento della feedback literacy degli studenti, in particolare nella sua dimensione peculiare di *taking action*, intesa come capacità di tradurre le informazioni provenienti dal feedback in azioni di miglioramento effettivo e trasformativo.

La feedback literacy dello studente infatti indica le conoscenze, le capacità e le disposizioni necessarie per dare un senso alle informazioni ricevute dai processi di feedback e utilizzarle per migliorare le strategie di lavoro o di apprendimento (Carless & Boud, 2018). Si tratta di una postura complessa, di cui fanno parte quattro aspetti di competenza tra loro interconnessi: saper apprezzare il feedback, formulare giudizi, gestire le emozioni e trasformare il feedback in azione.

Sulla base di tale framework, lo studio qui illustrato è stato condotto su un gruppo numericamente contenuto di studenti, al fine di privilegiare un'analisi in profondità dei significati da essi elaborati. Il dispositivo metodologico ha previsto l'impiego di artefatti di natura differenziata, tra cui domande a intenzionalità metacognitiva, somministrate in fasi distinte del percorso. La domanda metacognitiva viene qui intesa come dispositivo generativo del feedback, in grado di sollecitare processi di rielaborazione e azione, orientando implicitamente lo studente verso un uso attivo delle informazioni ricevute, in linea con quanto propongono Carless e Boud (2018, p. 2) relativamente a "make sense of information and use it to enhance work or learning strategies".

2. Descrizione del percorso e metodologia di analisi

La ricerca è stata realizzata nell'ambito di due insegnamenti dell'area pedagogica attivati presso due differenti Corsi di Laurea. I partecipanti erano 62 studenti, suddivisi in due gruppi di pari numerosità: 31 iscritti al Corso di Laurea in Scienze Pedagogiche dell'Università "G. d'Annunzio" di Chieti-Pescara e 31 iscritti al Corso di Laurea Magistrale in Matematica, curriculum didattico, dell'Università di Bologna.

Il disegno della ricerca si articolava in più fasi (Tab. 1) e si è svolto nell'arco di un mese: gli studenti hanno lavorato sia individualmente sia in piccoli gruppi per progettare un percorso educativo su un tema specifico di loro scelta, successivamente confrontato con un progetto analogo generato da ChatGPT.

Fasi dello studio esplorativo	**Descrizione**
1. Progettazione del percorso educativo	Progettazione, in piccolo gruppo, di un percorso educativo o disciplinare rivolto a uno specifico target.
2. Generazione di un percorso educativo da parte di ChatGPT	Realizzazione dello stesso compito da parte di ChatGPT, seguendo le medesime istruzioni fornite agli studenti.
3. Analisi del prodotto generato da ChatGPT: riflessione individuale	Compilazione di un questionario (Q1) al fine di: a) valutare il prodotto generato da ChatGPT; b) confrontarlo con il prodotto elaborato dal gruppo.
4. Analisi del prodotto generato da ChatGPT: riflessione di gruppo	Compilazione, da parte del gruppo, di una "scheda di confronto" fornita dai docenti, con l'obiettivo di: a) valutare il prodotto generato da ChatGPT; b) confrontarlo con il prodotto del gruppo; c) identificare punti di forza e criticità del prodotto generato da ChatGPT.
5. Analisi, da parte di ChatGPT, del progetto pedagogico creato dagli studenti	ChatGPT corregge e valuta il compito elaborato da ciascun gruppo sulla base di un prompt che richiede un feedback accurato in termini di punti di forza e aspetti da migliorare.
6. Riflessione di gruppo ed eventuale revisione del proprio testo	Discussione di gruppo sul feedback ricevuto da ChatGPT e compilazione di una "scheda strutturata" predisposta dai docenti.
7. Riflessione individuale ed eventuale revisione del proprio testo	Compilazione del questionario (Q2) per esprimere riflessioni metacognitive e attivare potenziali processi trasformativi.
8. Riorganizzazione delle conoscenze	Lezione congiunta dei docenti sull'Intelligenza Artificiale e sull'uso di ChatGPT nei contesti educativi.

Tab. 1 – Fasi dello studio esplorativo

Nelle prime fasi dell'attività (fase 1, 2, 3, 4) è stato chiesto loro di valutare la qualità del lavoro prodotto da ChatGPT: la finalità era quella di esplorare la possibilità di attivare il processo di feedback negli studenti partendo dal confronto tra l'elaborato prodotto dal gruppo e quello, su consegna identica, prodotto da ChatGPT. Questa fase è stata guidata dai docenti in modo da direzionare gli studenti verso quelle operazioni di osservazione, riflessione e modifica che rappresentano il cuore del mettere in azione il feedback: gli studenti hanno confrontato i due artefatti cercando di esplicitare le principali differenze e di comprendere come queste potessero orientare un eventuale miglioramento del

proprio elaborato. Nelle fasi successive (5, 6, 7) ChatGPT è stato impiegato per valutare i progetti elaborati dagli studenti, fornendo un feedback che i gruppi hanno analizzato e discusso al fine di affinare il proprio lavoro. La finalità era quella di esplorare la possibilità di attivare il processo di feedback negli studenti partendo, in questo caso, dalla valutazione prodotta dal chatbot, in vista di sollecitare gli studenti a osservare criticamente il proprio elaborato, a riflettere su ciò che poteva essere rivisto e a individuare possibili modifiche operative, sostenendoli nel passaggio dalla comprensione del commento ricevuto alla sua traduzione in azioni progettuali concrete. I risultati di questo processo sono stati utilizzati per capire le modalità attraverso cui l'IA generativa può essere integrata nei processi di apprendimento, con l'obiettivo di potenziare la students' agency nei contesti formativi, aiutando in particolare quelle pratiche di analisi, confronto e revisione che costituiscono la base del *taking action* all'interno della feedback literacy.

I dati, ampi e diversificati, sono stati raccolti tramite due questionari misti (Likert e domande aperte) somministrati agli studenti in momenti differenti del percorso formativo: un primo questionario (Q1), proposto all'inizio della seconda settimana, e un secondo questionario (Q2), proposto all'inizio della quarta settimana, in modo da intercettare fasi diverse del processo di elaborazione e rielaborazione dei feedback. Q1 è stato compilato da 61 studenti (45 F e 16 M, età media 23,2 anni), mentre Q2 da 57 studenti (42 F e 15 M, età media 23,4 anni).

Ai fini del presente contributo, l'attenzione si è concentrata sulle domande in grado di cogliere due dimensioni della feedback literacy: la prima dimensione, di natura *autovalutativa* e *riflessiva*, riguarda la capacità degli studenti di confrontare criticamente artefatti differenti, uno prodotto dal gruppo e uno generato dall'IA. Essa è stata indagata chiedendo ai partecipanti di valutare il prodotto elaborato da ChatGPT e di metterlo a confronto con il proprio (Q1). La seconda dimensione della feedback literacy, quella prettamente *trasformativa*, è legata alla capacità degli studenti di interpretare il feedback, sia esso derivante dal confronto tra artefatti (Q1) o fornito direttamente dal chatbot (Q2), e di tradurlo in azioni di miglioramento concrete ed efficaci (Tab. 2).

Le risposte aperte sono state analizzate secondo due modalità: la domanda Q1a) attraverso la *reflexive thematic analysis* (Braun & Clarke, 2019), adottando un approccio induttivo che consente di far emergere significati latenti, idee e concettualizzazioni a partire dai dati stessi; la codifica è stata rivista e concordata dalle tre autrici; le domande Q1b) e Q2c), invece, sono state ancorate a un orientamento deduttivo, coerente con la *directed content analysis* (Mayring, 2000), che ha consentito di interpretare i temi emersi alla luce di specifici costrutti teo-

rici sul feedback e sulla revisione. Sono state inoltre condotte analisi di frequenza dei tag, seguendo i principi dell'analisi qualitativa del contenuto (Krippendorff, 2018) e un'analisi delle co-occorrenze tra i tag, evidenziando alcuni pattern di concentrazione dei dati.

3. Risultati e discussione

I dati raccolti ci consentono di rilevare, attraverso la riflessione degli studenti, come si strutturano alcuni degli aspetti-chiave della feedback literacy e relativi soprattutto alla messa in atto dell'elaborazione e della riflessione sul feedback a fini trasformativi. Nella tabella 2 le domande sono connesse al tipo di feedback fornito dall'AIG, uno basato sul confronto e uno sul commento, e le relative posture esplicitate dagli studenti grazie alle differenti dimensioni del feedback (Carless & Boud, 2018).

Utilizzo di agenti AIG	Domanda	Elementi/modalità di feedback	Attivazione di posture	Dimensione del Feedback
Feedback come confronto	**Q1 domanda a)** Spiega le principali differenze tra l'elaborato del gruppo e l'elaborato prodotto da ChatGPT	Internal feedback	Conoscenza che lo studente genera quando confronta la propria idea con un'informazione di riferimento.	Dimensione autovalutativa e riflessiva
Feedback come confronto	**Q1 domanda b)** Dopo il confronto, modificheresti il tuo elaborato? In che modo?	Comprensione delle strategie di uptake	Come agire per migliorare.	Dimensione trasformativa
Feedback come commento e revisione	**Q2 domanda c)** Dopo il feedback dato dal ChatGPT, modificheresti il tuo elaborato? In che modo?	Apprezzamento del feedback, rendere il feedback attivo e generativo (taking action)	Cosa modificare? Come agire per migliorare?	Dimensione trasformativa

Tab. 2 – Mappatura delle domande analizzate in relazione alle dimensioni del feedback

3.1 Comparative feedback

L'analisi quantitativa delle risposte mostra che il confronto tra l'elaborato del gruppo e quello prodotto dall'AI ha attivato una riflessione ampia e multilivello

sulle principali dimensioni della competenza progettuale. Le categorie più ricorrenti risultano *mediazione didattica* (64,5%) e *strutturazione del percorso* (59,7%), seguite da *valutazione* (45,2%) e *finalità formative* (41,9%), con una presenza significativa anche di *differenziazione* (29%) e *processi cognitivi* (27,4%). Le frequenti co-occorrenze tra le categorie evidenziano una riflessione di tipo sistemico, in cui obiettivi, organizzazione, mediazione, valutazione e inclusione vengono letti in modo integrato. Dal punto di vista della feedback literacy, il comparative feedback ha sostenuto la capacità degli studenti di interpretare criticamente le informazioni di ritorno e di orientare la regolazione della progettazione. L'AI si configura così come un artefatto mediatore che, grazie alla regia didattica del docente, fornisce affordance per l'attivazione dell'inner feedback e di processi metacognitivi e autoregolativi orientati al miglioramento.

La scelta di strutturare il meccanismo di confronto deriva proprio dagli studi sul tema: Gentner e Kurtz (2006) sostengono che non è sufficiente la mera disamina di due artefatti, mentre il confronto ponderato e attento consente di andare oltre le differenze superficiali e mettere in evidenza le relazioni strutturali che vi sono tra i due prodotti. "Prompts, for example, asking students to identify similarities across the items being compared, increase the depth of relational encoding and the likelihood of learning transfer" (Nicol, 2021, p. 762). Inoltre, richiedere una spiegazione scritta in cui devono sostenere la validità delle loro spiegazioni rispetto agli elementi che hanno deciso di prendere in considerazione implementa la profondità del confronto (Edwards et al., 2019), produce un apprendimento più concreto e situato e trasferibile a nuovi contesti. Il processo verso una strutturazione di inner feedback (Nicol, 2019; 2021) si configura proprio dall'interazione con informazioni desunte da fonti esterne.

Wood (2022) evidenzia la necessità di costruire ambienti in cui sollecitare l'attività di confronto non solo per migliorare la comprensione di cosa sarebbe stato più opportuno fare, ma soprattutto per consentire agli studenti di mettere in atto strategie di *uptake*, ovvero di messa in atto, in pratica del feedback stesso.

Ciò rappresenta il cuore della feedback literacy, che si realizza proprio nella sua portata trasformativa dei saperi, dei meccanismi di apprendimento, delle strategie operative degli studenti (Laici, 2021).

3.2 Il feedback trasformativo

In questo senso sono state esplorate le risposte degli studenti sulle ipotesi di trasformazione del proprio compito ("Modificheresti il tuo elaborato? In che modo?"): una prima richiesta dopo l'attività di confronto, una seconda richiesta

dopo la seconda attività, che prevedeva l'erogazione del feedback sull'elaborato direttamente da parte di ChatGPT. Il dato più evidente è l'aumento dell'intenzione a modificare tra il primo e il secondo processo di feedback (i "no" diminuiscono del 17%): ciò fa supporre che un feedback più classico, erogato in forma di commento, sia maggiormente trasformativo in quanto invita direttamente lo studente all'azione indicando cosa e come modificare. Altrettanto interessante, tuttavia, è notare che molti degli studenti che hanno risposto "no" alla domanda "Intendi modificare il tuo compito dopo il feedback?", nel momento in cui si sono confrontati con la risposta aperta, che invitava a motivare e spiegare, sono tornati sui propri passi, facendo ipotesi di modifica (Tab. 3).

	dopo la prima attività (confronto)	**risp. no ma modifica**	**dopo la seconda attività (commento)**	**risp. no ma modifica**
sì	2 (3,28%)		4 (7,14%)	
no	29 (47,54%)	7 (24,14%)	17 (30,36%)	8 (47,06%)
in parte	30 (49,18%)		35 (62,50%)	

Tab. 3 –Confronto tra l'azione trasformativa del primo e del secondo processo di feedback

In tale dinamica si può vedere il potere del feedback, e nello specifico il potere che l'accompagnamento al feedback di cui parlano Wood (2022) e Nicol (2020) ha nei confronti dell'attivazione dello studente e dunque dell'alfabetizzazione al feedback: il ritorno autoregolativo e migliorativo diventa imprescindibile se sostenuto da adeguate domande di riflessione.

Per capire in che modo l'azione dello studente si reifica in concreto sono state identificate nelle risposte degli studenti volte alla modifica del compito quattro modalità di intervento e revisione. Le categorie sono formulate a partire da una ricognizione della letteratura sul *feedback uptake* (Carless & Boud, 2018; Nicol, 2019) e sul feedback applicato all'ambito dei *revision studies*, cercando di identificare le posture ristrutturative che si sono sviluppate a partire dal feedback ricevuto, cioè revisioni che modificano la struttura e il senso e non semplici correzioni superficiali (Bouwer & Dirkx, 2023). Le categorie identificate sono:

– *Aggiunte*: integrazione di nuovi contenuti o risorse per colmare lacune e ampliare l'artefatto: primo feedback 27, secondo feedback 20.
– *Rafforzamenti*: miglioramento di elementi già presenti per aumentarne coerenza, precisione e completezza senza modificarne la struttura: primo feedback 11, secondo feedback 21.

– *Riduzioni/eliminazioni*: rimozione di parti incoerenti o poco significative per aumentare l'efficacia del compito: primo feedback 0, secondo feedback 0.
– *Riorganizzazioni*: ristrutturazione profonda di sequenze e connessioni per ripensare l'impianto complessivo del lavoro: primo feedback 16, secondo feedback 18.

Il quadro delineato mostra che il primo feedback, quello generato dal confronto, sollecita gli studenti più negli aspetti correttivi che in quelli ristrutturativi, mentre il secondo direziona gli interventi di miglioramento verso la dimensione dell'approfondimento. Il feedback classico, in forma di commento diretto ed esplicito è più facile da accogliere e facilita una revisione più profonda poiché guida lo studente proattivamente. Nessuno studente agisce con riduzioni o eliminazioni di parti del compito: non vi sono evidenze sufficienti per spiegare tale assenza, che può essere determinata dalla difficoltà di chi si approccia alla progettazione nel prendere decisioni legate a logiche complesse quali la gerarchizzazione, la ridondanza o la deviazione (Capolla et al., 2024), che guidano le scelte rispetto a quali siano le parti importanti e imprescindibili e quelle non coerenti, ripetitive o inefficaci negli artefatti progettuali.

3.3 Il feedback riflessivo

L'altro tipo di analisi che è stata condotta è quella sul livello di riflessione attivato dal feedback. Le risposte sono state rilette e filtrate tramite una scala di livelli di riflessività, fondata su una sintesi dei principali modelli teorici relativi alla riflessione sulle pratiche: il livello superficiale riprende le forme di *non-reflection* di Kember et al. (2008), in cui lo studente si limita a modifiche operative senza reale comprensione. Il livello descrittivo corrisponde alla *descriptive reflection* (Hatton & Smith, 1995), nella quale si esplicita un primo grado di consapevolezza concettuale. Il livello critico si richiama alla *dialogic reflection* (Hatton, Smith, 1995) e alla *reflection* di Kember et al. (2008), in cui l'individuo rielabora il proprio agire individuando problemi, alternative e implicazioni. Il *livello metacognitivo*, infine, fa riferimento sia alla riflessività di Schön (1983), sia alla transformative reflection di Mezirow (1991), indicando una rielaborazione che trasforma gli schemi di pensiero e accresce la consapevolezza progettuale.

Livello di riflessione	Tipo di azione	% di risposte	Esempi
Superficiale	aggiunte operative, modifiche minime	*60%*	*"Forse si potrebbe aggiungere qualche dettaglio…" "Inserirei alcuni esercizi o problemi specifici"*
Descrittiva	migliorare chiarezza, coerenza, fasi	*22%*	*"Renderei più chiara la fase di valutazione" "Riorganizzerei la spiegazione perché più coerente"*
Critica	riconoscimento di un limite del progetto	*12,2%*	*"Mi sono resa conto che non avevamo previsto una vera valutazione formativa"*
Metacognitiva	riflessione sul proprio modo di progettare	*4,8%*	*"Il confronto mi ha fatto riflettere sul mio modo di progettare la lezione"*

Tab. 4 – Modifiche al compito dopo il feedback come confronto: livello di riflessione

Livello di riflessione	Tipo di azione	% di risposte	Esempi
Superficiale	modifiche operative	*24,4%*	*"Aggiungerei almeno un altro esperimento pratico…" "Modificherei il compito aggiungendo materiali come ad esempio libri…"*
Descrittiva	comprensione concettuale	*31,6%*	*"Modificherei le tempistiche in modo da avere un lavoro più specifico…" "Punterei a rendere la nostra proposta più ordinata e specifica"*
Critica	individuazione di limiti	*22%*	*"Sicuramente la parte relativa all'approfondimento dei contenuti, integrando informazioni teoriche oltre alla pratica." "L'equilibrio tra teoria e pratica… cancellando a priori anche ciò che di buono c'era."*
Metacognitiva	riflessione trasformativa	*22%*	*"Inizialmente ero contrario all'utilizzo della simulazione… ma l'osservazione del bot… mi ha fatto ricredere." "Ci ha dato spunti per riflettere su cosa volevamo veramente fare nella parte 1 dell'attività…"*

Tab. 5 – Modifiche al compito dopo il feedback come revisione: livello di riflessione

I dati mostrano un chiaro sviluppo nella profondità riflessiva tra il primo feedback (Tab. 4) e il secondo feedback (Tab. 5). Dopo il secondo feedback le risposte puramente superficiali calano drasticamente, mentre le risposte descrittive aumentano leggermente riscontrando inoltre più attenzione alla struttura generale del compito che non ad aggiunte sporadiche. Le risposte critiche e metacognitive raddoppiano quasi in percentuale, indicando un *uptake* più profondo del feedback ricevuto e una maggiore capacità di autovalutazione e trasformazione professionale.

I dati sono limitati e il tempo della sperimentazione troppo esiguo per poter affermare che la feedback literacy, implementando le strategie di feedback e differenziandole, possa piano piano diventare più strutturale e caratterizzata da un *taking action* più riflessivo e meno procedurale o operativo. Tuttavia, il processo di incorporazione della postura può dirsi avviato, poiché gli studenti aumentano la loro familiarità con il feedback e imparano a gestire cicli di feedback provenienti da fonti differenti e generati da agenti non esclusivamente umani, come nel caso del feedback del docente o del peer feedback. La costruzione di ambienti di feedback ibrido può diventare una strategia replicabile, una palestra di sperimentazione in cui gli studenti imparano a mettersi in gioco e a dialogare criticamente con gli agenti artificiali.

4. Conclusioni

L'esperienza descritta mostra come l'AI entri a pieno titolo negli ecosistemi formativi, i quali diventano spazi *more-than-human*, entro cui è possibile attivare processi riflessivi a patto di non pretendere di sostituire nel processo di mediazione e di gestione del feedback l'agente umano con l'agente artificiale. Il feedback che emerge in questi spazi è sempre più ibrido, frutto dell'interazione tra docente, pari e agenti artificiali, ridefinendo il concetto stesso di peer feedback in chiave ibrida, estesa e reticolare. L'AI si conferma strumento utile, che genera possibilità ma non significati, domande ma non risposte: l'attribuzione di senso resta responsabilità epistemica condivisa tra docente e studente. In questo quadro, i contesti formativi diventano vere e proprie palestre di competenze interattive, in cui si apprendono posture dialogiche, critiche e autoregolative. Il fine ultimo non è il feedback in sé, ma la costruzione della competenza, dell'*inner feedback* come disposizione stabile e consapevole alla trasformazione.

Riferimenti bibliografici

Bouwer R., & Dirkx K. (2023). The eye-mind of processing written feedback: Unraveling how students read and use feedback for revision. *Learning and Instruction, 85*, 101745.

Braun V., & Clarke V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589-597.

Capolla L.M., Giannandrea L., Gratani F., Pentucci M., & Rossi P.G. (2024). Rethinking and formalizing initial teacher training on learning design for and in uncertainty. *Frontiers in Education*, 9, 1268936. DOI: 10.3389/feduc.2024.1268936.

Carless D. (2019). Feedback loops and the longer-term: towards feedback spirals. *Assessment & Evaluation in Higher Education*, *44*(5), 705-714.
Carless D., & Boud D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315-1325.
Floridi L. (2025). *La differenza fondamentale. Artificial Agency: una nuova filosofia dell'intelligenza artificiale*. Milano: Mondadori.
Deleuze G., & Guattari F. (1980). *Mille plateaux.* (Vol. 4). Paris: éd. de Minuit.
Edwards, B. M., Cameron, D., King, G., & McPherson, A. C. (2019). How Students without Special Needs Perceive Social Inclusion of Children with Physical Impairments in Mainstream Schools: A Scoping Review. *International Journal of Disability, Development and Education, 66*(3), 298–324.
Gentner, D., & Kurtz, K. J. (2006). Relations, objects, and the composition of analogies. *Cognitive science, 30*(4), 609-642.
Hatton, N., & Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education, 11*(1), 33-49.
Hwang, G.-J.,& Chang, C.-Y. (2021). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments, 30*(10), 1-14.
Kember, D., McKay, J., Sinclair, K., & Wong, F. K. Y. (2008). A four category scheme for coding and assessing the level of reflection in written work. *Assessment & Evaluation in Higher Education, 33*(4), 369-379.
Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. London: Sage.
Laici, C. (2021). *Il feedback come pratica trasformativa nella didattica universitaria*. Milano: FrancoAngeli.
Latour, B. (1994). Pragmatogonies: A Mythical Account of How Humans and Nonhumans Swap Properties: A Mythical Account of How Humans and Nonhumans Swap Properties. *American Behavioral Scientist, 37*(6), 791-808.
Mezirow, J. (1991). *Transformative Dimensions of Adult Learning.* San Francisco: Jossey-Bass.
Morris, R., Perry, T. W., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: a systematic review. *Review of Education, 9*(3), 1-26.
Nicol, D. (2019). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 44*(5), 770-786.
Nicol, D. (2021). The power of internal feedback: exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756-778.
Schön, D. (1983), Il professionista riflessivo. Roma: Dedalo.
Schmulian, A., & Coetzee, S. A. (2019). Students' experience of team assessment with immediate feedback in a large accounting class. *Assessment & Evaluation in Higher Education, 44*(4), 516-532.
Wood, J. (2022). Making peer feedback work: the contribution of technology-mediated dialogic peer feedback to feedback uptake and literacy. *Assessment & Evaluation in Higher Education, 47*(3), 327-346.

II.24
Efficacia del feedback docente e IA: un'analisi delle percezioni nel contesto accademico[1]

Livia Petti
Università telematica Pegaso, livia.petti@unipegaso.it
Marta De Angelis
Università del Molise, marta.deangelis@unimol.it
Andrea Garavaglia
Università telematica Pegaso, andrea.garavaglia@unipegaso.it

Abstract

Il feedback riveste un ruolo cruciale nell'esperienza universitaria e l'IA genera interrogativi sul ruolo che possa svolgere nell'elaborazione del feedback. Questo studio si propone di analizzare le percezioni degli studenti rispetto a due differenti tipologie di feedback (quello elaborato dal docente e quello generato dall'IA) con l'obiettivo di valutare eventuali differenze, nonché di indagare quale dei due sia percepito come più efficace per sostenere i processi di apprendimento. 62 soggetti volontari hanno ricevuto entrambi i tipi di feedback che sono stati percepiti come chiari e utili; il 51,6% li ha ritenuti complementari. L'IA può supportare i docenti, ma la sua efficacia dipende dalla qualità del prompt, da rubriche strutturate e dall'ancoraggio ai materiali didattici per garantire coerenza e ridurre distorsioni.

Parole chiave: feedback; intelligenza artificiale; università.

Feedback plays a crucial role in the university experience, and AI raises questions about the role it can play in feedback processing. This study aims to analyze students' perceptions of two different types of feedback (that provided by the teacher and that generated by AI) with the aim of evaluating any differences, as well as investigating which of the two is perceived as more effective in supporting learning processes. Sixty-two volunteers received both types of feedback, which were

1 Il presente contributo è frutto del lavoro congiunto degli autori. Nello specifico sono attribuibili a Livia Petti i paragrafi 1 e 2; a Marta De Angelis il paragrafo 3. Il paragrafo 4 è stato scritto da Marta De Angelis e Andrea Garavaglia. Infine, il paragrafo 5 è stato redatto congiuntamente dagli autori.

perceived as clear and useful; 51.6% considered them complementary. AI can support teachers, but its effectiveness depends on the quality of the prompt, structured rubrics, and anchoring to teaching materials to ensure consistency and reduce bias.

Keywords: feedback; artificial intelligence; higher education.

1. Introduzione

Il feedback è fondamentale per l'apprendimento, ma è difficile offrirlo in modo tempestivo e personalizzato in ambito universitario, soprattutto quando ci sono molti studenti. Esso riveste un ruolo cruciale nell'esperienza universitaria dei corsisti, influenzando in modo significativo le percezioni che i discenti sviluppano nei confronti delle situazioni valutative in ambito accademico (Bartram & Bailey, 2010). Comprendere tali percezioni si rivela particolarmente rilevante, poiché processi come la valutazione e il feedback sono spesso vissuti dagli studenti come momenti critici e ad alto impatto nel loro percorso formativo (Brown, 2014; Doria et al., 2025; Nicol, 2010). Negli ultimi anni, l'introduzione dell'Intelligenza artificiale (IA) nei contesti educativi ha suscitato un crescente interesse, sollevando interrogativi sul ruolo che essa possa svolgere nell'elaborazione e nella restituzione del feedback agli studenti. I sistemi avanzati di IA come i Large Language Models (LLM) sono oggi in grado di comprendere, elaborare e generare il linguaggio umano, tali azioni possono contribuire ad aiutare il docente a fornire feedback personalizzati in modo rapido e mirato alle specifiche attività (Fidan & Gencel, 2022; Lim et al., 2020).

In questo quadro, il contributo si propone di analizzare le percezioni degli studenti rispetto a due differenti tipologie di feedback – quello elaborato dal docente e quello generato dall'IA – con l'obiettivo di valutare eventuali differenze, nonché di indagare quale dei due sia percepito come più efficace per sostenere i processi di apprendimento.

2. Background teorico

Il feedback è un elemento essenziale nel processo di insegnamento-apprendimento: se formulato in modo adeguato e fornito tempestivamente contribuisce significativamente alla comprensione degli studenti (Hattie & Timperley, 2007; Henderson et al., 2021; Wisniewski et al., 2020). Per essere efficace, il feedback

deve essere chiaro, orientato al compito (Ryan et al., 2019), fornire indicazioni pratiche su come proseguire e motivare gli studenti a utilizzarlo per migliorare le proprie prestazioni (Shute, 2008); è altresì essenziale che venga fornito in tempi utili, affinché gli studenti possano metterlo in pratica (Hounsell, 2007).

Un feedback di alta qualità non si limita a comunicare ai discenti la correttezza delle loro performance, ma offre indicazioni più profonde: supporta nell'affrontare il compito (processo) e incoraggia gli studenti nella gestione autonoma del proprio apprendimento (apprendimento autoregolato) (Hattie & Timperley, 2007). I discenti possono, infatti, utilizzare il feedback ricevuto per modificare le attività già svolte e per migliorare le performance future (Carless, 2019; Hounsell, 2007).

Pur essendo in grado di generare feedback di alta qualità, i docenti possono trarre vantaggio dall'uso dell'IA, che semplifica notevolmente questo compito, specialmente nel contesto universitario e in aule in cui è presente un gran numero di studenti. Infatti, il feedback automatizzato fornito dall'IA può accelerare e ampliare il processo di erogazione del riscontro (Cavalcanti et al., 2021; Keuning et al., 2019). In particolare, l'IA generativa, e gli LLM, stanno emergendo come strumenti efficaci per l'automazione del feedback in ambito educativo.

In ambienti stabili, l'IA può offrire prestazioni adeguate: creando un prompt ben strutturato e fornendo al sistema materiali pertinenti (come attività, rubriche di valutazione ed esempi) si aumenta notevolmente la probabilità di avere dal sistema risposte significative. Gli algoritmi complessi tendono infatti a funzionare meglio in contesti ben definiti, dove sono disponibili dati, contrariamente all'intelligenza umana che ha sviluppato capacità di gestione dell'incertezza, indipendentemente dalla quantità di informazioni a disposizione. Automatizzare il feedback consente al docente di risparmiare tempo, permettendogli di dedicare maggiore attenzione ad altre attività importanti come preparare le lezioni, supportare gli studenti e svolgere attività di ricerca (Alshater, 2022).

Alcuni studi mostrano come gli studenti apprezzino l'immediatezza e la specificità del feedback generato dall'IA (Lee & Moore, 2024), sebbene persistano preoccupazioni riguardo all'affidabilità del feedback fornito dall'IA rispetto a quello umano (Nazaretsky et al., 2024).

I docenti mostrano un atteggiamento generalmente positivo verso il feedback prodotto dall'IA, pur riconoscendo alcune limitazioni, quali la potenziale ridondanza e la presenza di errori (Tzirides et al., 2024). L'approccio considerato più efficace è quello che integra la personalizzazione dell'IA con la supervisione umana, dove l'efficienza tecnologica e la profondità pedagogica si combinano in modo proficuo (Mirzaei & Khayer, 2025).

3. Metodologia e strumenti

L'obiettivo dell'indagine è quello di esplorare in che modo gli studenti percepiscono l'efficacia di due differenti tipologie di feedback nel contesto universitario: tale confronto risulta particolarmente rilevante alla luce del crescente impiego dell'IA nei processi valutativi e metacognitivi.

Sulla base della letteratura e degli obiettivi dello studio, sono state formulate due domande di ricerca:

– RQ1: in che modo gli studenti percepiscono l'efficacia, la credibilità e l'impatto sul proprio apprendimento del feedback prodotto dal docente rispetto a quello generato dall'IA?
– RQ2: con adeguati accorgimenti nella progettazione (es. uso di rubriche valutative, prompt mirati e materiali di riferimento), l'IA è in grado di generare un feedback che presenti un livello di similarità semantica comparabile a quello del feedback redatto dal docente?

L'indagine è stata svolta nell'a.a 2024/25 e ha coinvolto gli studenti di tre insegnamenti universitari dell'Università di Milano (*Progettazione didattica e valutazione*, 9 cfu, 60 ore) e dell'Università del Molise (*Didattica e metodologie interattive*, 8 cfu, 48 ore; *Valutazione degli apprendimenti*, 6 cfu, 36 ore). Il campione è stato ottenuto attraverso un campionamento per convenienza, coinvolgendo gli studenti dei corsi partecipanti che hanno volontariamente aderito all'indagine (n = 87).

In ciascuno dei tre insegnamenti è stata proposta un'attività diversa, specifica per il relativo corso. Per ogni attività svolta, ciascuno studente ha ricevuto due feedback distinti: il primo, redatto dal docente, e il secondo generato tramite l'IA. Per la generazione dei feedback tramite IA è stato utilizzato *Notebook LM* (*Google*): prima della produzione definitiva sono stati effettuati vari test al fine di definire un prompt efficace e coerente con il modello di feedback e con la rubrica valutativa utilizzata dai docenti per valutare i prodotti. Sono stati quindi caricati, all'interno dell'applicativo, un file contenente i materiali di studio del corso, la rubrica valutativa da utilizzare per la generazione dei feedback e le consegne individuali degli studenti. Il prompt finale richiedeva a *Notebook LM* di produrre un feedback argomentato, della lunghezza massima di 750 battute, basandosi sul modello fornito e sui materiali caricati. Tale prompt è stato applicato ai compiti degli studenti per ottenere un feedback strutturato e coerente con i criteri valutativi. I due feedback sono stati successivamente restituiti in forma anonima, privi di qualsiasi indicazione della fonte, al fine di evitare effetti

di pregiudizio legati alla percezione di autorevolezza o familiarità con il docente o con l'IA.

La raccolta dei dati è stata effettuata attraverso un questionario online anonimo, somministrato agli studenti contestualmente alla ricezione dei due feedback. Lo strumento è stato costruito dagli autori sulla base della letteratura e di questionari già validati sulla percezione dell'efficacia del feedback (Caro-Cahilig & Cabanero, 2024; Jellicoe & Forsythe, 2019; Lipnevich et al., 2021; Strijbos et al., 2010; 2021). Esso è composto da una scala Likert a sei punti (1= fortemente in disaccordo; 6= fortemente d'accordo) e comprende diverse sezioni volte a indagare aspetti complementari di entrambi i feedback: caratteristiche specifiche, credibilità percepita della fonte, risposta emotiva e accettazione, cambiamenti comportamentali e impatto complessivo sull'apprendimento; una sezione è dedicata alla comparazione tra i due feedback, chiedendo agli studenti quale ritengano più efficace e la relativa motivazione. Degli 87 studenti che hanno aderito allo studio, 62 hanno compilato effettivamente il questionario, costituendo il campione di rispondenti su cui si basa l'analisi delle risposte. Sono inoltre stati raccolti dati socio-anagrafici (corso, anno di studio, età, genere) utili per analisi descrittive e confronti tra sottogruppi.

Parallelamente alla raccolta dei dati relativi alla percezione degli studenti, è stata condotta un'analisi di similarità semantica[2] dei due feedback tramite il software SEMPL-IT[3], con l'obiettivo di descrivere il grado di convergenza tra i contenuti prodotti dal docente e quelli generati dall'IA. L'integrazione di dati quantitativi, qualitativi e linguistico-computazionali ha consentito di ottenere un quadro articolato e metodologicamente solido delle percezioni degli studenti e delle caratteristiche dei due feedback.

2 Per similarità semantica si intende il grado di vicinanza di significato tra due testi: essa misura quanto due formulazioni diverse esprimano concetti equivalenti o sovrapponibili, indipendentemente dalle parole utilizzate o dall'ordine in cui compaiono (Chandrasekaran & Mago, 2021). Tale misura non si limita al confronto superficiale delle parole, ma considera le relazioni concettuali, il contenuto informativo condiviso e il modo in cui i testi veicolano il significato complessivo.

3 Il software SEMPL-IT nasce come applicativo open access specializzato nella semplificazione automatica di testi amministrativi, ma integra anche funzionalità avanzate di analisi linguistica utili per gli scopi della ricerca. In particolare, è in grado di confrontare la similarità semantica tra due testi. https://sempl-it.unimol.it/

4. Risultati e discussione

Per quanto riguarda i feedback generati dal docente e dall'IA, essi mostrano un alto livello medio di similarità semantica (82%), con una bassa variabilità tra i casi (DS= 0,034), suggerendo che i due tipi di feedback risultano mediamente molto coerenti tra loro nel contenuto.

Anche l'intervallo di confidenza al 95% [0,813 – 0,827] conferma la stabilità della stima, indicando che i valori di similarità si distribuiscono entro un margine molto ridotto attorno alla media. L'istogramma (Fig. 1) mostra infatti una concentrazione delle frequenze tra 0,80 e 0,86, mentre il range complessivo dei valori (0,75-0,92) risulta contenuto e privo di outlier rilevanti, suggerendo un elevato grado di coerenza tra i feedback analizzati.

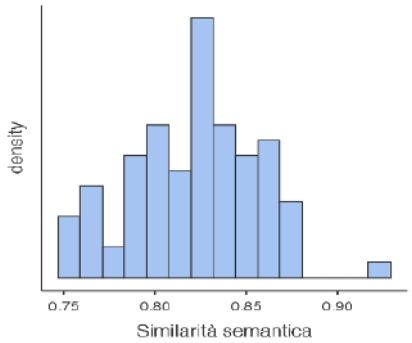


Fig.1 – Distribuzione dei valori di similarità semantica tra feedback del docente e feedback generati dall'IA

Anche i valori medi per corso mostrano un'elevata convergenza semantica in tutte le condizioni, con variazioni contenute tra gli insegnamenti (Tab.1), confermando una sostanziale stabilità del comportamento semantico dell'IA nei diversi contesti didattici.

Insegnamento	Media	DS
Metodologie didattiche interattive	0,82	0,0383
Progettazione didattica e valutazione	0,81	0,0350
Valutazione degli apprendimenti	0,83	0,0250

Tab.1 – Similarità semantica per insegnamento (media e deviazione standard)

In merito invece ai dati ricavati dal questionario, i rispondenti (n= 62) si distribuiscono per insegnamento nel modo seguente: il 50% frequenta *Didattica e metodologie interattive*, il 31% *Valutazione degli apprendimenti* e il 19% *Progettazione didattica e valutazione*. La maggior parte degli studenti si colloca al secondo anno di corso (77%), seguita dal terzo anno (15%) e da una quota più contenuta di studenti magistrali (8%). L'età media è di 23 anni (DS= 5) e il campione risulta prevalentemente femminile, con il 92% di studentesse e l'8% di studenti.

L'analisi comparativa tra il feedback fornito dal docente e quello generato dall'IA restituisce un quadro di marcata convergenza qualitativa. I dati contenuti nella tabella 2 evidenziano come gli studenti abbiano valutato entrambe le fonti in modo estremamente positivo, con medie aggregate che si attestano sistematicamente nella fascia alta della scala (valori superiori a 5 su 6).

Sezione	Feedback docente Media (DS)	Feedback IA Media (DS)
1. Caratteristiche del feedback	5,39 (1,05)	5,54 (0,92)
2. Credibilità della fonte	5,49 (0,86)	5,53 (0,99)
3. Accettazione emotiva	5,43 (0,93)	5,48 (0,95)
4. Cambiamento comportamentale e intenzioni	5,11 (1,19)	5,30 (1,07)
5. Impatto complessivo	5,41(0,97)	5,53 (0,91)

Tab. 2 – Statistiche descrittive aggregate per sezioni del questionario

Le differenze rilevate, pur indicando una lieve preferenza per l'IA, risultano minime e descrivono una sostanziale sovrapponibilità dell'efficacia didattica. Nello specifico, per quanto riguarda le *caratteristiche del feedb*ack e la sua credibilità (sezioni 1 e 2), l'IA si dimostra un interlocutore percepito come altrettanto autorevole dell'umano. La fiducia riposta nello strumento digitale (M= 5,58) è statisticamente indistinguibile da quella verso il docente (M= 5,53), con uno scarto trascurabile (≤ 0,05). Entrambe le modalità condividono inoltre la stessa criticità: la tempestività è l'unico parametro che registra punteggi inferiori, suggerendo che la percezione del "tempo utile" rimane una sfida comune indipendentemente dalla fonte.

Sul piano dell'*accettazione emotiva* (sezione 3), i risultati smentiscono l'ipotesi di una "freddezza" algoritmica. Al contrario, l'IA sembra ridurre l'ansia da giudizio, generando livelli di sicurezza nelle proprie capacità (M= 5,44) ed emozioni positive (M= 5,31) superiori a quelli registrati con il feedback umano (rispetti-

vamente 5,32 e 5,16). Il dato trova conferma anche nelle risposte aperte: "il Feedback 2 (IA), oltre a indicare chiaramente cosa è stato fatto bene e cosa manca, usa un tono incoraggiante che aiuta a sentirsi più sicuri e motivati a migliorare. Questo tipo di feedback non si limita a spiegare gli errori, ma sostiene anche la fiducia nelle proprie capacità, rendendo più facile mettere in pratica i suggerimenti ricevuti" (Cod. 10). Questo elemento aumenta l'autoefficacia degli studenti incoraggiando la gestione autonoma del proprio apprendimento (Hattie & Timperley, 2007). Tuttavia, il docente mantiene un primato specifico nella capacità di stimolare la volizione: la disposizione al miglioramento risulta infatti più alta nel feedback prodotto dall'insegnante (M= 5,65 contro 5,52 dell'IA), suggerendo che la relazione umana rimane un fattore chiave per la motivazione all'impegno. Infine, nell'ambito del *cambiamento comportamentale* e dell'*impatto sull'apprendimento* (sezioni 4 e 5), si osserva la maggiore, seppur contenuta, variazione. L'IA appare leggermente più incisiva nel modificare le abitudini di studio (M= 4,89 vs 4,56) e nel promuovere la ricerca di risorse aggiuntive. Tuttavia, la percezione complessiva dell'utilità per il raggiungimento degli obiettivi formativi rimane elevata per entrambi i gruppi (M>5,4), confermando che l'introduzione dell'IA non sostituisce, ma eguaglia la validità del supporto docente.

Particolarmente interessante è poi la sezione di *comparazione tra le fonti*, in cui gli studenti sono chiamati a esprimere una preferenza tra il feedback del docente e quello generato dall'IA. Oltre la metà dei rispondenti (51,6%) considera i due feedback complementari, ritenendo che insieme offrano una restituzione più completa. Un ulteriore 17,7% giudica i feedback sostanzialmente equivalenti, mentre le preferenze univoche si distribuiscono in modo equilibrato: 14,5% a favore del docente e 16,1% a favore dell'IA. Questa distribuzione risulta coerente con l'elevata similarità semantica riscontrata tra i due feedback (M=0,82), suggerendo che, per molti studenti, le due fonti veicolano contenuti concettualmente allineati e pertanto percepiti come ugualmente utili e credibili.

Per quanto riguarda le motivazioni che giustificano quale feedback viene preferito, si riporta di seguito una sintesi delle analisi. Coloro che considerano i feedback complementari hanno sottolineato che i due feedback, pur simili nel contenuto, offrivano prospettive diverse ma necessarie (es. più spunti pratici, integrazione chiarimenti e consigli per migliorare). Invece, chi considera i feedback simili ha messo in evidenza che la sostanza è la stessa anche quando la forma è differente. Ancora, coloro che preferiscono il Feedback 1, proposto dal docente, spiegano che ne apprezzano molto l'utilità, la struttura, la schematicità e la capacità di andare "dritti al punto". Infine, chi preferisce il Feedback 2, quello generato dall'IA, spiega che risulta più chiaro e con maggiori dettagli. In questo

gruppo emerge anche un piccolo sottogruppo che dichiara di percepire il feedback maggiormente empatico, paradossalmente maggiormente "umano" o incoraggiante rispetto al primo.

5. Conclusioni

I risultati hanno dimostrato una notevole convergenza semantica tra le due tipologie di feedback, suggerendo che l'IA, se opportunamente guidata da prompt ben strutturati, rubriche di valutazione dettagliate e ancoraggio a materiali didattici specifici, è in grado di produrre riscontri di elevata qualità contenutistica, non dissimili da quelli elaborati dal docente. Sul piano della formazione docente, i risultati evidenziano la necessità di includere l'uso dell'IA all'interno delle competenze professionali legate alla valutazione formativa. In particolare, la capacità di progettare rubriche condivise, di definire criteri valutativi espliciti e di supervisionare criticamente il feedback generato dall'IA, si configura come un ambito specifico di sviluppo professionale, necessario affinché questi strumenti possano essere integrati in modo consapevole nei processi valutativi universitari. La strategia più efficace, infatti, non risiede nella sostituzione, ma nell'adozione di un modello ibrido che integri l'efficienza e la specificità dell'IA con la profondità pedagogica e la dimensione relazionale del feedback umano. Investire in percorsi formativi di questa tipologia può contribuire alla diffusione di pratiche valutative più sostenibili, trasparenti e orientate allo sviluppo dell'apprendimento autoregolato degli studenti.

Riferimenti bibliografici

Alshater, M. (2022). *Exploring the Role of Artificial Intelligence in Enhancing Academic Performance: A Case Study of ChatGPT*. Available at SSRN: https://ssrn.com/abstract=4312358 or http://dx.doi.org/10.2139/ssrn.4312358

Bartram, B., & Bailey, C. (2010). Assessment preferences: A comparison of UK/international students at an English university. *Research in Post-Compulsory Education*, *15*(2), 177-187.

Brown, S. (2014). *Learning, teaching and assessment in higher education: Global perspectives.* London: Palgrave Macmillan.

Carless, D. (2019). Feedback loops and the longer-term: Towards feedback spirals. *Assessment & Evaluation in Higher Education*, 44(5), 705–714.

Caro-Cahilig. R. & Cabanero, (2024). Teachers' Feedback Techniques Questionnaire Development And Validation Of The Instrument For The Study Teachers' Feedback

Techniques Its Influence On Learners' Performance. *International Journal of Academic Multidisciplinary Research*, *8*(4), 168-170.
Cavalcanti, A.P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.S., Gaševi , D. & Mello, R.F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence* 2, 100027.
Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—a survey. *Acm Computing Surveys* (Csur), *54*(2), 1-37.
Doria, B., Foschi, L. C., Raffaghelli, J. E. & Grion, V. (2025). University students' perceptions and experiences of teacher, peer and automatic feedback. *Italian Journal of Educational Research*, 34, 135 149.
Fidan, M., & Gencel, N. (2022). Supporting the instructional videos with Chatbot and peer feedback mechanisms in online learning: The effects on learning performance and intrinsic motivation. *Journal of Educational Computing Research*, *60*(7).
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
Henderson, M., Ryan, T., Boud, D., Dawson, P., Phillips, M., Molloy, E., Mahoney, P. (2021). The usefulness of feedback. *Active Learning in Higher Education* 22, 229–243.
Hounsell, D. (2007). Chapter 8: Towards more sustainable feedback to students. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the longer term* (pp. 101–113). UK: Routledge.
Jellicoe, M., & Forsythe, A. (2019). The development and validation of the feedback in learning scale (FLS). In *Frontiers in Education* (Vol. 4, p. 84). Frontiers Media SA.
Keuning, H., Jeuring, J. & Heeren, B. (2019). A Systematic Literature Review of Automated Feedback Generation for Programming Exercises. *ACM Transactions on Computing Education* 19, 1–43.
Lee, S. S., & Moore, R. L. (2024). Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review. *Online Learning*, 28(3).
Lim, L.-A., Dawson, S., Gaševi , D., Joksimovi , S., Pardo, A., Fudge, A., & Gentili, S. (2020). Students' sense-making of personalised feedback based on learning analytics. *Australasian Journal of Educational Technology*, *36*(6), 15-33.
Lipnevich, A. A., Gjicali, K., Asil, M., & Smith, J. K. (2021). Development of a measure of receptivity to instructional feedback and examination of its links to personality. *Personality and Individual Differences*, 169, 110086.
Mirzaei, S., & Khayer, B. (2025). Enhancing feedback personalisation with AI-generated analytics: A narrative review. *Learning Letters*, 5, 50.
Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J. & Käser, T. (2024). AI or Human? Evaluating Student Feedback Perceptions in Higher Education. In: Ferreira Mello, R., Rummel, N., Jivet, I., Pishtari, G., Ruipérez Valiente, J.A. (eds) *Technology Enhanced Learning for Inclusive and Equitable Quality Education. EC-TEL 2024. Lecture Notes in Computer Science*, vol 15159. Springer, Cham.
Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, *35*(5), 501-517.

Ryan, T., Henderson, M., Phillips, M. (2019). Feedback modes matter: Comparing student perceptions of digital and non-digital feedback modes in higher education. *British Journal of Educational Technology,* 50(3), 1507–1523.

Shute, V.J. (2008). Focus on Formative Feedback. *Review of Educational Research* 78, 153–189.

Strijbos, J. W., Pat-El, R. J., & Narciss, S. (2010). Validation of a (peer) feedback perceptions questionnaire. In *Proceedings of the International Conference on Networked Learning* (Vol. 7, pp. 378-386).

Strijbos, J. W., Pat-El, R., & Narciss, S. (2021). Structural validity and invariance of the feedback perceptions questionnaire. *Studies in Educational Evaluation*, 68, 100980.

Tzirides, A.O., Zapata, G.C., Bolger, P., Cope, B., Kalantzis, M., & Searsmith, D. (2024). Exploring Instructors' Views on Fine-Tuned Generative AI Feedback in Higher Education. *International Journal on E-Learning*, 23(3), 319-334.

Wisniewski, B., Zierer, K. & Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology,* 10.

II.25
Designing intelligent assessment systems: an MCP-based framework for UDL - oriented feedback[1]

Umberto Barbieri
Università Telematica Pegaso, umberto.barbieri@unipegaso.it
Lia Daniela Sasanelli
Università Telematica Pegaso, liadaniela.sasanelli@unipegaso.it
Raffaele Di Fuccio
Università Telematica Pegaso, raffaele.difuccio@unipegaso.it

Abstract

The integration of Artificial Intelligence into assessment processes promises to revolutionize feedback and learning personalization. Nevertheless, large-scale adoption is hindered by significant challenges: the stochastic nature of Large Language Models (LLM) undermines their reliability; tool heterogeneity produces non-standardized results; teacher workflows remain inefficient. This contribution proposes an architecture based on the Model Context Protocol (MCP) to overcome these limitations, ensuring consistency, control, and transparency. The framework implements the UDL Guidelines 3.0 for generating "action-oriented feedback", translating the pedagogical theoretical principles into operational tools. Three integrated modules are presented—ontological organization of sources, UDL-oriented instructional design, feedback generation—accompanied by an evaluative rubric and an observation grid. The framework can be used in all subject areas, from primary school through to university, to support teachers in delivering truly formative, fair and inclusive assessments.

Keywords: Artificial Intelligence; Formative Assessment; Feedback; Universal Design for Learning; Model Context Protocol.

L'integrazione dell'Intelligenza Artificiale nei processi valutativi promette di rivoluzionare feedback e personalizzazione dell'apprendimento. L'adozione su larga scala è tuttavia ostacolata da sfide significative: la natura stocastica dei Large Language Models ne compromette l'affidabilità; l'eterogeneità degli strumenti pro-

1 This contribution is the result of the joint work of the authors. However, paragraphs 3 and 5 are attributed to Umberto Barbieri; paragraphs 2 and 4 to Lia Daniela Sasanelli; paragraphs 1 and 6 to Raffaele Di Fuccio.

duce risultati non standardizzati; i flussi di lavoro per i docenti risultano inefficienti. Il presente contributo propone un'architettura basata sul Model Context Protocol (MCP) per superare queste limitazioni, garantendo consistenza, controllo e trasparenza. Il framework implementa le linee guida UDL 3.0 per la generazione di "feedback orientati all'azione", traducendo i principi teorici di Bandura, Schunk, Hattie e Dweck in strumenti operativi. Vengono presentati tre moduli integrati – organizzazione ontologica delle fonti, progettazione didattica UDL-oriented, generazione di feedback – corredati da una rubrica valutativa e una griglia di osservazione. Il framework è utilizzabile dal primo ciclo d'istruzione fino all' università, in tutte le aree disciplinari, supportando i docenti in una valutazione realmente formativa, equa ed inclusiva.

Parole chiave: Intelligenza Artificiale; Valutazione formativa; Feedback; Universal Design for Learning; Model Context Protocol.

1. Introduction

The adoption of Large Language Models (LLM) in educational contexts has preceded any systematic reflection on their use. ChatGPT, Claude, Gemini: tools with documented capabilities—generation of instructional materials, automated formative feedback, adaptive personalization (Hardi et al., 2025; Swacha et al., 2025). But their effective implementation presupposes competencies that teachers, typically, do not possess. Recent literature confirms this asymmetry. Knoth et al. (2024) show that output quality depends directly on prompt quality, and that prompt engineering constitutes a specific competency—not a natural extension of disciplinary expertise. Walter (2024) identifies three skills necessary for effective AI use in education: AI literacy, prompt engineering, critical thinking. None falls within standard teacher training. A comparison can clarify the nature of the problem. No one asks a teacher to learn HTML to view a webpage, nor to master TCP/IP to send an email. Generative AI does not offer, in its current form, an analogous separation. Output effectiveness depends on input quality to a degree that requires—in effect—implicit programming competencies. The paradox is evident. The very tools designed to make instruction more personalized and inclusive prove accessible only to those with advanced technical competencies. The teacher who cannot formulate a structured prompt obtains generic responses, sometimes hallucinations, pedagogically useless content (Gao et al., 2025). The tool promised to save time, and at the end it wastes it. The alternative—training all teachers in prompt engineering—is not realistic. The speed of evolution of these systems renders any

training obsolete within months (Celik et al., 2022). It is unreasonable to expect teachers to become AI engineers, just as we do not expect them to become web developers. The solution lies not in training users in complexity, but in building infrastructures that hide it. The *Model Context Protocol* (MCP), introduced by Anthropic in 2024, offers a technical foundation for this separation: an open standard that normalizes communication between AI applications and external systems, exposing predefined resources, functions, and workflows to any compatible client. This contribution proposes using this protocol to build an intelligent assessment architecture based on the principles of Universal Design for Learning (CAST, 2024; Meyer et al., 2014). The goal is not to teach teachers to use AI but designing systems that allow them to achieve professional results without having to master the underlying technology.

2. Theoretical Framework: UDL Feedback

Bandura's social cognitive theory (1986, 1977, 1997) places self-efficacy perception at the center as a determining factor in motivation, persistence, and resilience: students who believe in their capabilities invest greater effort in tasks, persist in the face of difficulties, and select challenges proportionate to their possibilities, as demonstrated by multifactorial analyses (Bandura et al., 1996). This perception is strengthened primarily through four channels directly influenceable by calibrated feedback: direct mastery experiences, observation of others' successes (vicarious experiences), positive verbal persuasion, and management of negative emotional states. Schunk (1982, 1984, 1984) integrates this framework with the concept of attributional feedback, experimentally demonstrating that attributing results to effort and strategies produces statistically significant increases in self-efficacy (d = 0.68) and performance, clearly surpassing feedback based on innate gifts. The distinction is operational, not merely theoretical: “You worked with dedication and your score increased by 20%” produces measurably different effects from “You have a talent for this subject.” Hattie (2012), in his meta-analysis of over 1,200 studies, places feedback among high-impact influences (d = 0.73), emphasizing that maximum effectiveness is achieved when it informs students on how to improve the learning process, not just the final product, and follows a logical progression: recognition of effort, introduction of specific strategies, and celebration of progress (Schunk & Cox, 1986; Schunk & Zimmerman, 1997). In parallel, Dweck and Leggett (1988) and Dweck (2013) distinguish the fixed mindset—which attributes success to innate talent, favoring challenge avoidance and post-failure collapse—from the growth mindset—

which views intelligence as expandable through effort, promoting resilience and continuous growth. Longitudinal studies on 373 adolescents quantify average gains of 0.4 standard deviations in academic outcomes (Blackwell et al., 2007), while controlled experiments reveal that praising effort preserves persistence and intrinsic motivation, unlike praise for intelligence which erodes performance by 10% (Mueller & Dweck, 1998). Trinchero and Calvani (2019) contextualize these principles within the Italian didactic model, stating that effective teaching is based on formative feedback that supports cyclical self-correction toward shared objectives.

2.1 UDL Guideline 8.5

The theoretical framework just outlined has been fully incorporated into the Universal Design for Learning framework, within the Guidelines 3.0, through didactic consideration 8.5 "*Provide feedback that is oriented toward action*" (CAST, 2024). The Guidelines emphasize that student assessment processes are most productive when feedback is relevant, constructive, accessible, consistent, and timely (Mitwally et al., 2026).The type of feedback is also fundamental in helping students maintain motivation and engagement, essential for learning. Action-oriented feedback offers specific comments on how to progress and act on work to reach the learning goal. This typology emphasizes the role of effort and practice, rather than intelligence or innate ability, as an important factor in guiding students toward successful long-term mental habits and learning practices and in promoting resilience. The UDL Guidelines 3.0 (CAST, 2024) provide explicit guidance for producing action-oriented feedback, articulated across six operational dimensions described below.

– *Perseverance, self-efficacy, and self-awareness.* Feedback should promote persistence, self-efficacy perception (Bandura, 1997), and metacognitive self-awareness (Zimmerman, 2000), encouraging the use of specific supports and strategies when facing difficulties.
– *Effort, improvement, and goals.* Feedback should prioritize recognition of invested effort, observable progress, and goal achievement, avoiding performance comparisons. Stating goals to be reached guides students, directs effort in the correct direction, and allows them to monitor their performance relative to shared targets.
– *Frequency, timeliness, and specificity.* Feedback should be frequent, temporally proximate to the action, and highly specific about learning processes. Recent

meta-analyses (Hattie, 2012; Bangert-Downs et al., 1991) attest to its high impact on self-regulatory cycles (Schunk & Zimmerman, 1997).

– *Substantive and informative content.* Feedback should favor personal and informative content, excluding comparative or competitive judgments that undermine self-efficacy (Dweck & Leggett, 1988). Hattie (2012) classifies generic praise as low-impact compared to process-oriented feedback, as it provides no concrete guidance on how to improve.
– *Metacognitive reflection on patterns of strength and weakness.* Feedback should model how to analyze one's recurring learning patterns, transforming challenges and strengths into transferable strategies through sequential steps: identify patterns, translate into strategies, activate self-regulated cycle (Zimmerman, 2000).
– *Risk-taking and alternative perspectives.* Feedback should encourage "controlled cognitive risk" and diverse perspectives, normalizing "productive struggle" (Dweck, 2013) with scalable scaffolds for all students. The central idea is to transform feedback into a tool that not only corrects errors but stimulates students to venture further.

3. Methodology: Architecture of an MCP Server for UDL Assessment

The *Model Context Protocol* (Fig. 1) defines three primitives through which a server exposes functionalities to compatible AI clients: resources (informational resources – Module 1), tools (executable functions – Module 2), prompts (structured workflows – Module 3). This architecture shifts control from the language model to the system designer—in educational contexts, to the institution (Lameras & Arnab, 2022). The server described here implements these primitives in three modules oriented toward Universal Design for Learning: ontological organization of content, assisted instructional design, generation of action-oriented feedback.

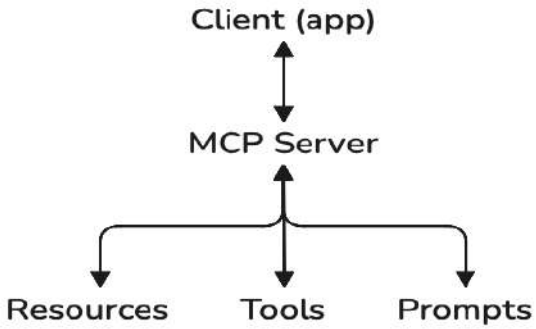


Fig. 1 – MCP server architecture showing bidirectional client-server communication and access to Resources, Tools, and Prompts

3.1 Module 1: Ontological Organization of Sources

Personal notes, slides, manuals, audio and video files: teaching practice accumulates resources in an unstructured corpus, difficult to navigate, impossible to query systematically. The first module exposes a resource (knowledge-base) and a tool (build_ontology) that transform these heterogeneous materials into a queryable ontology. The process operates in three phases.

The system indexes uploaded documents, extracting key concepts, hierarchical relationships, and cross-connections. It then generates a structured representation (Xu et al., 2025): interconnected notes, each with explicit metadata on prerequisites, learning objectives, and associated UDL checkpoints. The teacher can then query, modify, and enrich this structure through the AI client, without intervening on the original files. The output is not an index but represents a navigable map that makes explicit relationships that otherwise remains implicit: which concepts precede which, where typical difficulties concentrate, which resources cover which objectives (Fig. 2). This transparency constitutes the prerequisite for subsequent modules.

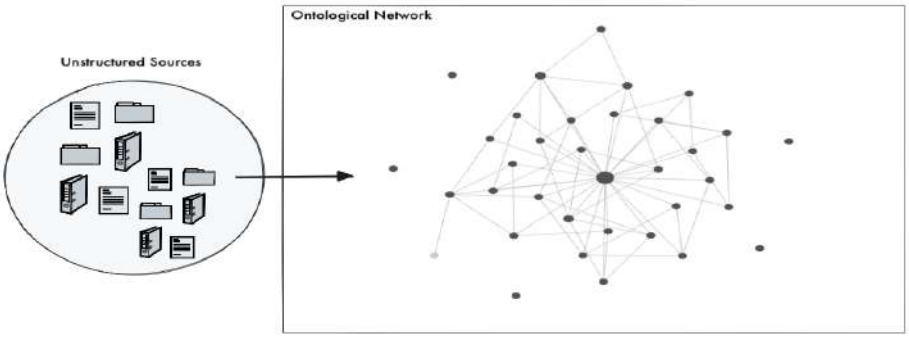


Fig. 2 – Transformation of unstructured educational resources
into a queryable ontological network

3.2 Module 2: UDL-Oriented Instructional Design

The second module assists in designing materials conforming to the UDL Guidelines 3.0 (CAST, 2024). It exposes a set of tools that operate on the ontology generated by the first module: generate_udl_metadata, check_accessibility, suggest_alternatives.

The first tool is *generate_udl_metadata*, it analyzes an instructional unit and produces a structured report according to the three UDL principles: engagement, representation, action & expression. For each principle, it identifies checkpoints already satisfied and those absent, suggesting specific integrations. If the unit presents exclusively textual content, the system flags the missing coverage of checkpoint 1.1 ("*Offer ways of customizing the display of information*") and proposes concrete alternatives: audio version, visual schema, integrated glossary.

The second one is *check_accessibility* and it verifies technical compliance of materials according to WCAG 2.1—color contrast, typographic readability, presence of alternative text for images. Finally, the *suggest_alternatives* generates variants of the same content calibrated to different learning profiles, operationalizing the UDL principle of flexibility. These tools do not replace pedagogical judgment. They provide a systematic analysis that the teacher can accept, modify, or ignore. The value lies in completeness: no checkpoint overlooked due to oversight, gaps visible before delivery.

3.3 Module 3: Action-Oriented Feedback

The third module constitutes the operational core of the system. It translates into executable functions checkpoint 8.5 of the UDL Guidelines: "*Provide feedback that is oriented toward action*" (CAST, 2024). The theoretical base integrates the four frameworks discussed in the previous section: Bandura's self-efficacy theory, Schunk's attributional feedback, Hattie's classification of feedback impact, Dweck's distinction between fixed and growth mindsets.

The module translates these principles into a main tool (*generate_udl_feedback*) and two supporting resources: an evaluative rubric, an observation grid. The tool receives as input student work, evaluation criteria, and—optionally—historical data on previous performance extracted from the ontology. The output is feedback structured according to eight operational dimensions:

- *Anchoring to objective*: what the student is learning, what they did, what effect it produced.
- *Contextual relevance:* why the comment is relevant now, to which proximate goal it connects.
- *Effort-based attribution:* recognition of effort and strategies with quantifiable data, no reference to innate gifts.
- *Suggested strategy:* identify a concrete and feasible action, not a generic criticism
- *Next step:* proposes a specific action within a defined time frame, indicating an expected product.
- *Reflective stimulus:* suggests questions that activate metacognitive and reflective processes
- *Normalization of difficulty:* valuing attempts, errors reframed as useful information.
- *Formal accessibility:* clear language, segmented structure, asset-based tone

3.4 Module Integration

The three modules do not operate in isolation. The ontology built by the first feeds the other two: instructional design draws from the content map to verify coverage and coherence; the feedback system accesses historical data to contextualize individual progress. Integration produces a multiplier effect. The teacher who uses the system for an entire school year progressively has access to an increasingly rich knowledge base—more precise feedback, more targeted designs.

From a technical standpoint, the server exposes these functionalities through standardized endpoints. Any compatible client—Claude, ChatGPT with plugins, custom applications—can connect and operate according to the same rules, with the same data, producing comparable outputs. Standardization is the condition that makes reliability possible.

4. Results: Tools for Assessment Practice

The architecture described produces three operational outputs intended to support daily assessment practice of the teachers: a synoptic table for feedback construction, an evaluative rubric for quality verification, an observation grid for longitudinal monitoring.

4.1 Synoptic Table for UDL Feedback Construction

Table 1 operationally synthesizes the steps necessary to construct feedback conforming to UDL principles. For each dimension, it identifies the key question the teacher must ask, the theoretical alignment with reference frameworks, and concrete operational guidance.

Key Question	Theoretical Alignment	Operational Guidance
What do I want the student to understand from my feedback?	Direct mastery experiences (Bandura, 1997)	Always write: 1) "You are working on..." (goal); 2) "You did X..." (behavior); 3) "This allowed you to..." (effect)
Why is this feedback important for them right now?	Process feedback and "where to go next" (Bandura, 1997; Hattie, 2012)	Add "This feedback helps you because...". Connect to proximate and controllable goal
How can I recognize effort attributionally?	Effort-based attributional feedback (Schunk, 1982, 1984; Mueller & Dweck, 1998)	"You dedicated [time/attempts] and this led to [measurable improvement]". Avoid "You're smart"
What specific strategy can I suggest?	Concrete strategy for perception of control (Schunk, 1983; Bandura, 1997)	One clear strategy at a time: "Next time, try to...". Associate with visible support
What next step can I propose?	Agency and self-regulation (Hattie, 2012; Schunk & Zimmerman, 1997)	1) Precise action; 2) Defined timeframe; 3) Expected product. Avoid "study more"
How can I support metacognitive reflection?	Self-regulation cycles and growth mindset (Zimmerman, 2000; Dweck, 2017)	Guide questions: "What helped you most?", "Where do you struggle?", "What strategy will you reuse?"

Key Question	Theoretical Alignment	Operational Guidance
How do I promote the courage to take risks?	Growth mindset (Dweck & Leggett, 1988; Dweck, 2013)	Name the positive risk: "You made a courageous attempt...". Connect error to useful information
In what format do I make feedback accessible?	Cognitive load reduction (Okolo, 1992; Calvani, 2014)	Short sentences, bullet structure. Asset-based tone. Avoid overload

Tab. 1 – Operational Guide for UDL Feedback Construction

4.2 UDL Feedback Evaluative Rubric

The evaluative rubric (*Tab.* 2) allows teachers to verify the quality of generated feedback—whether from the AI system or manually written—before sending it to students. The four-level scale (1 = Absent, 2 = Partial, 3 = Good, 4 = Excellent) operationalizes each criterion with observable behavioral descriptors.

Criterion	1 – Absent	2 – Partial	3 – Good	4 – Excellent
1. Relevance and clarity of objective	Does not indicate the objective; uses generic phrases ("good", "bad")	References the activity, objective implicit or unclear	Names the objective but does not connect it to student action	Clearly indicates objective and connects action to progress
2. Attribution to effort and strategies	Privileges talent/intelligence; does not mention effort	Cites effort vaguely, mixed with judgments about abilities	Recognizes effort and strategy but without concrete data	Specifically recognizes effort and strategy with quantifiable data
3. Action orientation	Does not indicate what to do next	Generic request without concrete steps	Proposes 1-2 actions but without times or expected products	Proposes specific actions with times and expected product
4. Specificity and timeliness	Rare, only at end of course; generic or very delayed	Infrequent or bundled, with significant delay	Moderately timely, fairly specific guidance	Frequent, close to activity, highly specific
5. Promotion of self-efficacy	Language that undermines self-efficacy; no reflective questions	Neutral/positive tone but reflection absent	Invites simple reflection	Models reflection: progress, strengths/weaknesses, future strategies
6. Encouragement of perseverance	Emphasizes error or failure; discourages attempts	Mixes recognition and criticism without valuing attempt	Recognizes attempt and invites retry	Normalizes difficulty, values cognitive risk
7. Accessibility and form	Complex language, excessively long text	Clear but dense language; single format	Clear language, bullet points; at least two modalities	Simple, segmented language; multiple modalities; asset-based tone

Tab. 2 – UDL Feedback Evaluative Rubric

4.3 UDL Feedback Observation Grid

The observation grid (*Tab.* 3) is designed to support self-monitoring of assessment practice over time. For each criterion, it provides an operational description guiding the observer—the teacher themselves or a supervisor—in score assignment. Systematic use of the grid enables identification of recurring patterns in one's feedback practice and areas for improvement.

Criterion	Operational Description	Code (1-4)
1. Clarity of objective and message	Feedback explicates what the student is learning and connects what they did to the result obtained. Assess how clearly the three elements (objective, behavior, effect) are present.	☐1 ☐2 ☐3 ☐4
2. "Here and now" relevance	Feedback explains why the comment matters now and to which proximate goal it connects, avoiding vague references or only final grades.	☐1 ☐2 ☐3 ☐4
3. Attributional recognition	Feedback attributes results to effort and strategies, not innate gifts. Check for "you're talented/not talented" versus phrases quantifying effort and progress.	☐1 ☐2 ☐3 ☐4
4. Suggested strategy	Feedback proposes at least one concrete and practicable strategy, possibly referencing tools/models, rather than merely stating what is missing.	☐1 ☐2 ☐3 ☐4
5. Concrete next step	Feedback clearly indicates the next step: precise action, defined timeframe, and expected product, enabling a task feedback revision cycle.	☐1 ☐2 ☐3 ☐4
6. Promotion of reflection	Feedback contains at least one question or request for reflection and/or invites the student to identify strengths, weaknesses, and future strategies.	☐1 ☐2 ☐3 ☐4
7. Support for perseverance	Feedback values attempts and normalizes difficulty, presenting errors as useful information. Note if "courage to try" is mentioned.	☐1 ☐2 ☐3 ☐4
8. Accessibility and form	Feedback uses clear language, short sentences, bullet structure; avoids unnecessary jargon and deficit-based formulations. Note format adaptations when needed.	☐1 ☐2 ☐3 ☐4

Tab. 3 – UDL Feedback Observation Grid
Note: Coding scale: 1 = insufficient, 2 = basic, 3 = good, 4 = excellent

5. Discussion

The implications of the proposed architecture touch all thematic areas of education: didactics, ethics, training. Focusing on the didactic level, the framework enables scalable and consistent formative assessment. The standardization guaranteed by the MCP protocol ensures that any compatible AI client, connecting to the institutional server, operates according to the same rules and with the

same data. The teacher using Claude or ChatGPT obtains feedback structured according to the same UDL criteria, eliminating variability introduced by prompt quality. Consistency is no longer a function of the user's technical expertise. At the ethical level, the architecture increases transparency and accountability (Allison, 2025) AI operations are explicit: the teacher—or the institution—defines a priori which functions are available, which resources are accessible, which workflows are permitted. Control over potential algorithmic biases becomes possible because the system does not operate as a black box but as inspectable infrastructure. The assessment tools (rubric, grid) make output quality verifiable before it reaches the student. Finally, considering the training level, a new professional profile emerges. The teacher evolves from AI user to designer of intelligent assessment systems (Balaji et al., 2025). This transition does not require programming competencies—the MCP protocol provides the necessary abstraction—but pedagogical competencies applied to system design. Deciding which resources to include in the knowledge base, which UDL instructional considerations to monitor and which feedback dimensions to prioritise are pedagogical decisions, not technical ones. The framework returns to teachers the control that generative AI complexity seemed to subtract from them. The architecture also presents limitations that deserve explicit discussion. By design, the MCP server encapsulates all pedagogical complexity—prompt templates, UDL-aligned evaluation criteria, ontological structures, and workflow logic—so that the end user operates within a fully pre-configured environment requiring no prompt engineering or system design expertise. However, this advantage comes at the cost of an initial installation step that is less immediate than accessing a generalist chatbot through a simple web interface. The user must download or connect to the MCP server and integrate it with a compatible AI client (e.g., Claude, ChatGPT, or a custom application), a process whose complexity varies depending on the provider's connection procedures. Additionally, the framework presupposes institutions willing to invest in maintaining the infrastructure over time: updating resources as curricula evolve, refining prompt calibration based on usage feedback, and ensuring server availability.

6. Conclusions

This contribution has proposed an architecture based on the Model Context Protocol for formative assessment in educational contexts. Integration of the UDL Guidelines 3.0, translated into operational tools through the three de-

scribed modules, enables generation of action-oriented feedback conforming to the principles of Bandura, Schunk, Hattie, and Dweck without requiring prompt engineering competencies from teachers. The main factor is the definition of a tool that is anchored on the pedagogical theory and on the UDL guidelines leading to a system based on rooted assumptions. A consistent advantage lays on the technical standardization. In practical terms any compatible client operates according to the same rules produces pedagogical standardization: comparable, verifiable feedback, improvable over time. The evaluative rubric and observation grid provide tools for quality control and self-monitoring of practice. This is central in the development of the tool because guarantee a human-in-the-loop approach (Moscheira-Ray et al., 2023). The system is not immediate for any user but allow a strong improvement in opposition to a real knowledge in prompt engineering. The added value is on the possibility to use a system that is built on strong scientific evidence, and the user could consider to develop teaching actions inspired by the proposed system without a profound knowledge of the theories. At the current stage of development, the technical architecture is fully implemented: all three modules (ontological organization, UDL-oriented instructional design, and action-oriented feedback generation) are operational and integrated within a functional MCP server. The infrastructure, endpoints, resources, and tools described in Section 3 are not theoretical proposals but working components. What remains to be completed is the systematic validation of the prompt templates embedded in the system. The system will therefore be fully operational and available for institutional adoption once prompt validation is finalized. Future developments will include empirical validation of the framework in real school contexts, measurement of impact on learning outcomes and student self-efficacy, and extension of the system to peer assessment contexts. The final horizon is to provide teacher with system highly controlled in their development, allowing the teacher to focus their attention on the teaching/learning level, and on the personalization on the classroom and the students that represent the final target.

References

Allison, J. (2025). RAISE the Standard: A Framework for Transparent Reporting of Artificial Intelligence Studies in Education. *Journal of Educational Computing Research*, 07356331251377430. https://doi.org/10.1177/07356331251377430.

Balaji, V., Ogange, B. O., & Mays, T. J. (2025). Teacher in the Loop AI (TiL-AI): A Strategy for Empowering Educators in Developing Countries through OER Adap-

tation. *Journal of Learning for Development*, 12(2), 439–445. https://doi.org/-10.56059/jl4d.v12i2.1934
Bandura, A. (1977). *Social learning theory* (pp. viii, 247). Prentice-Hall.
Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.
Bandura, A. (1997). *Self-efficacy: The exercise of control* (pp. ix, 604). W H Freeman/Times Books/ Henry Holt & Co.
Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. *Child Development*, 67(3), 1206–1222.
Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research*, 61(2), 213-238. https://doi.org/10.3102/00346543061002213
Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit Theories of Intelligence Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention. *Child Development*, 78(1), 246–263. https://doi.org-/10.1111/j.1467-8624.2007.00995.x
Calvani, A. (2014). *Come fare una lezione efficace*. Roma: Carocci.
CAST. (2024). *Universal Design for Learning Guidelines version 3.0.* Wakefield, MA: Author. (n.d.).
Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The Promises and Challenges of Artificial Intelligence for Teachers: A Systematic Review of Research. *TechTrends*, 66(4), 616–630. https://doi.org/10.1007/s11528-022-00715-y.
Dweck, C. S. (2013). *Self-theories: Their Role in Motivation, Personality, and Development*. Psychology Press. https://doi.org/10.4324/9781315783048.
Dweck, C. S. (2017). From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological Review*, 124*(6),* 689–719. https://doi.org/10.1037/rev0000082
Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95(2), 256–273. https://doi.org/10.1037/0033-295X.95.2.256.
Gao, F., He, Y., Chen, Q., & Liu, F. (2025). Evaluating Psychological Competency via Chinese Q&A in Large Language Models. *Applied Sciences*, 15(16), 9089. https://doi.org/10.3390/app15169089.
Hardi, I., Nur, R., Ammade, S., Latifa, A., & Syawal. (2025). Effective Feedback Strategies in Online English Learning: A Systematic Literature Review. *Inspiring: English Education Journal*, 8(2), 201–221. https://doi.org/10.35905-/inspiring.v8i2.14523.
Hattie, J. (2012). *Visible Learning for Teachers: Maximizing Impact on Learning*. Routledge. https://doi.org/10.4324/9780203181522.
Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6, 100225. https://doi.org/10.1016/j.caeai.2024.100225
Lameras, P., & Arnab, S. (2022a). Power to the Teachers: An Exploratory Review on

Artificial Intelligence in Education. *Information*, 13(1), 14. https://doi.org/10.3390/info13010014.

Lameras, P., & Arnab, S. (2022b). Power to the Teachers: An Exploratory Review on Artificial Intelligence in Education. *Information,* 13(1), 14. https://doi.org/-10.3390/info13010014

Mitwally, M. A., Cheniti-Belcadhi, L., & Hadyaoui, A. (2026). Universal Design for Learning and AI-Enhanced Assessment: Developing a Comprehensive Ontology for Equitable Evaluation. In H. Zaky (Ed.), *Innovating Assessment and Evaluation in Higher Education: Inclusive Practices and Technological Advancements* (pp. 41-76). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-2130-1.ch002.

Modelcontextprotocol/modelcontextprotocol. (2025). *Model Context Protocol.* https://github.com/modelcontextprotocol/modelcontextprotocol (Original work published 2024)

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, *56*(4), 3005-3054.

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology*, 75(1), 33–52. https://doi.org/10.1037/0022-3514.75.1.33

Okolo, C. M. (1992). The effects of computer-based attribution retraining on the attributions, persistence, and mathematics computation of students with learning disabilities. *Journal of Learning Disabilities,* 25(5), 327-334. https://doi.org/10.1177/002221949202500509

Schunk, D. H. (1982). Effects of effort attributional feedback on children's perceived self-efficacy and achievement. *Journal of Educational Psychology,* 74(4), 548–556. https://doi.org/10.1037/0022-0663.74.4.548

Schunk, D. H. (1983). Ability versus effort attributional feedback: Differential effects on self-efficacy and achievement. *Journal of Educational Psychology,* 75(6), 848-856. https://doi.org/10.1037/0022-0663.75.6.848

Schunk, D. H. (1984). Sequential attributional feedback and children's achievement behaviors. *Journal of Educational Psychology*, 76(6), 1159–1169. https://doi.org/10.1037/0022-0663.76.6.1159

Schunk, D. H., & Cox, P. D. (1986). Strategy training and attributional feedback with learning disabled students. *Journal of Educational Psychology*, 78*(3), 201*–209. https://doi.org/10.1037/0022-0663.78.3.201

Schunk, D. H., & Zimmerman, B. J. (1997). Social origins of self-regulatory competence. *Educational Psychologist,* 32(4), 195–208 https://doi.org/-10.1207/s15326985ep3204_1

Swacha, J., & Gracel, M. (2025). Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Applied Sciences,* 15(8), 4234. https://doi.org/10.3390/app15084234.

Trinchero, R. Calvani, A. (2019). *Dieci falsi miti e dieci regole per insegnare bene*. Roma: Carocci.

Walter, Y. (2024). Embracing the future of Artificial Intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21(1), 15. https://doi.org/10.1186/s41239-024-00448-3

Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., & Zhang, Y. (2025). A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82-91. https://doi.org/10.1006/ceps.1999.1016

II.26
Il feedback e l'apprendimento. Cosa cambia nell'era dell'intelligenza artificiale

Simona Michelon
PhD student and Economics Teacher
Università degli studi di Modena
simo1miche@gmail.com

Abstract

La valutazione formativa rappresenta una sfida educativa. Sebbene l'intelligenza artificiale (IA) offra nuove opportunità per generare feedback tempestivi e personalizzati, il suo uso appare ben lontano dall'essere integrato nella didattica. Lo studio ha coinvolto circa 400 insegnanti ed è ispirato al *Conversational Framework* di Diana Laurillard. I dati sono raccolti attraverso un questionario strutturato.I dati sono analizzati mediante statistiche descrittive e inferenziali. I risultati rivelano una cultura pedagogica del feedback diffusa che non si traduce in modo uniforme nella pratica. L'adozione di feedback impliciti rappresenta la dimensione più fragile della competenza valutativa. L'analisi dei cluster ha identificato tre profili professionali che evidenziano la necessità di percorsi formativi differenziati. In tutti i profili, il supporto istituzionale è emerso determinante.

Parole chiave: feedback; conversational framework; intelligenza artificiale; apprendimento; insegnamento.

Formative assessment represents a significant educational challenge. Although artificial intelligence (AI) offers new opportunities to generate timely and personalised feedback, its use remains far from being fully integrated into teaching practices. The study involved approximately 400 teachers and is grounded in Diana Laurillard's *Conversational Framework*. Data were collected through a structured questionnaire and analysed using descriptive and inferential statistics.
The findings reveal a widespread pedagogical culture of feedback that is not consistently translated into practice. The adoption of implicit feedback emerges as the most fragile dimension of assessment competence. Cluster analysis identified three professional profiles, highlighting the need for differentiated training pathways. Across all profiles, institutional support proved to be a determining factor.

Keywords: feedback; conversational framework; artificial intelligence; learning; teaching.

1. Introduzione

La valutazione formativa rappresenta una delle leve più rilevanti per il miglioramento degli apprendimenti e per l'innovazione educativa, in particolare nell'attuale contesto di trasformazione digitale. All'interno di questo quadro, il feedback assume un ruolo centrale, ma non univoco: la letteratura mostra come il feedback possa essere inteso secondo prospettive differenti, ciascuna delle quali mette in luce specifiche funzioni pedagogiche e condizioni di efficacia.

Il presente studio si fonda su tre filoni teorici principali che, pur condividendo l'idea del feedback come elemento chiave dell'apprendimento, ne offrono letture complementari: il feedback come informazione regolativa (Hattie & Timperley), il feedback come processo dialogico e progettuale (Laurillard) e il feedback come competenza professionale e relazionale (Carless & Winstone). Queste prospettive costituiscono la base teorica su cui si innesta l'analisi delle convinzioni pedagogiche e delle pratiche valutative dei docenti italiani.

Nel modello proposto da Hattie e Timperley (2007), il feedback è concepito principalmente come informazione regolativa (*regulative feedback*), ovvero come un insieme di informazioni che aiutano lo studente a ridurre la distanza tra la prestazione attuale e l'obiettivo di apprendimento atteso. Il feedback è efficace nella misura in cui orienta il processo di apprendimento rispondendo a tre domande fondamentali: *Where am I going?*, *How am I going?*, *Where to next?*

Questa prospettiva distingue diversi livelli di feedback – sul compito, sul processo, sull'autoregolazione e sulla persona – evidenziando come l'impatto maggiore sull'apprendimento sia associato al feedback che agisce sui processi cognitivi e metacognitivi, piuttosto che su valutazioni meramente correttive o personali. In questa lettura, il feedback è uno strumento potente ma prevalentemente **esplicito**, intenzionale e osservabile, che assume la forma di commenti, indicazioni e orientamenti forniti dal docente.

Una seconda prospettiva fondamentale è offerta dal *Conversational Framework* di Laurillard (2012), che interpreta il feedback come parte integrante di un processo dialogico e ciclico di costruzione della conoscenza (*dialogic feedback*). In questo modello, l'apprendimento non è visto come una semplice ricezione di informazioni, ma come un continuo confronto tra le idee dello studente, le rappresentazioni del docente e l'azione sul compito.

Il feedback, in questa prospettiva, non è un evento puntuale né un momento finale della valutazione, ma il meccanismo strutturante dell'interazione educativa**.** Esso emerge sia in forma estrinseca (fornita dal docente o dal sistema) sia in forma intrinseca, quando lo studente confronta l'esito delle proprie azioni con gli obiettivi prefissati. Il valore pedagogico del feedback risiede quindi nella

sua capacità di sostenere cicli successivi di azione, riflessione e riformulazione, rendendolo un elemento centrale della progettazione didattica.

Più recentemente, Carless e Winstone (2023) hanno proposto una lettura del feedback centrata sulla feedback literacy del docente, intesa come un insieme di competenze professionali che rendono il feedback efficace, sostenibile e significativo nel tempo. In questa prospettiva, il feedback è concepito come una pratica relazionale e riflessiva (*relational feedback*), che non si esaurisce nella trasmissione di informazioni, ma richiede capacità di progettazione, giudizio valutativo, sensibilità relazionale e riflessività professionale.

La feedback literacy del docente include la capacità di progettare feedback coerenti con gli obiettivi di apprendimento, di costruire relazioni di fiducia che ne favoriscano l'accoglienza e di osservare come gli studenti utilizzano il feedback per adattare le proprie pratiche. Questa prospettiva mette in luce in particolare il ruolo del feedback implicito, ovvero di quei segnali, regolazioni e micro-interazioni che orientano l'apprendimento senza assumere forme esplicitamente dichiarative.

Nel loro insieme, queste tre prospettive delineano un quadro complesso del feedback educativo:

– come informazione regolativa (Hattie & Timperley),
– come processo dialogico e ciclico (Laurillard),
– come competenza professionale e relazionale (Carless & Winstone).

Il presente studio si colloca all'intersezione di questi approcci, assumendo che il feedback non sia solo una tecnica valutativa, ma una pratica pedagogica situata, che dipende dalle convinzioni dei docenti, dalle loro abitudini operative e dalle condizioni istituzionali in cui operano. In particolare, l'attenzione alla distinzione tra feedback esplicito e implicito consente di esplorare in che misura le diverse concezioni teoriche del feedback trovino riscontro nelle pratiche reali dei docenti.

2. Metodologia

La ricerca si inserisce all'interno di un più ampio percorso di ricerca-azione rivolto a docenti italiani di diversi ordini e gradi scolastici. L'obiettivo metodologico di questo contributo è duplice: da un lato, indagare il sistema di convinzioni pedagogiche dei docenti in relazione al feedback; dall'altro, analizzare le pratiche valutative abitualmente adottate, distinguendo tra feedback esplicito e feedback

implicito. Il disegno metodologico è stato costruito per garantire robustezza statistica, coerenza teorica e interpretabilità pedagogica.

Il campione è composto da docenti volontari iscritti al percorso di formazione. La distribuzione per ordine di scuola è eterogenea e comprende sia il primo sia il secondo ciclo di istruzione. Pur trattandosi di un campione non probabilistico, tale eterogeneità consente di esplorare in modo articolato le differenze nei profili professionali e nelle pratiche valutative.

I dati sono stati raccolti mediante un questionario strutturato composto esclusivamente da item a risposta chiusa. Le analisi previste includono statistiche descrittive, analisi di affidabilità, analisi fattoriale esplorativa e analisi dei cluster, in coerenza con gli obiettivi esplorativi dello studio.

Il questionario è stato progettato a partire dalla letteratura in particolare dal *Conversational Framework* di Laurillard (2012), che interpreta l'apprendimento come un processo ciclico e dialogico sostenuto dal feedback continuo. In questa prospettiva, il feedback non è concepito come un momento valutativo finale, ma come un meccanismo regolativo integrato nel processo di apprendimento, capace di orientare l'azione, la riflessione e la riorganizzazione concettuale degli studenti.

Coerentemente con tale cornice teorica, la survey è stata costruita per indagare diversi aspetti del rapporto tra docenti e feedback, con particolare attenzione al grado di conoscenza e consapevolezza del feedback come leva pedagogica; le credenze e gli atteggiamenti dei docenti rispetto al valore formativo del feedback; le pratiche valutative abituali, distinguendo tra forme di feedback esplicito e implicito; la percezione delle condizioni istituzionali che favoriscono o ostacolano una cultura del feedback; il ruolo percepito delle tecnologie digitali e dell'intelligenza artificiale nel supportare i processi di feedback e valutazione.

Le domande del questionario derivano direttamente dalla distinzione, proposta da Laurillard, tra feedback estrinseco e feedback intrinseco all'interno del ciclo dell'apprendimento. Tale distinzione è stata tradotta operativamente in item volti a rilevare sia pratiche osservabili e dichiarative (feedback esplicito), sia forme più tacite e situate di regolazione dell'apprendimento (feedback implicito), insieme alle convinzioni e agli atteggiamenti che ne orientano l'uso.

Nel suo insieme, il questionario mira a cogliere non solo ciò che i docenti dichiarano di fare, ma anche come interpretano il feedback all'interno del processo di apprendimento e in che misura percepiscono il contesto istituzionale e tecnologico come supportivo rispetto a tali pratiche.

La ricerca si inserisce all'interno di un più ampio percorso di ricerca-azione rivolto ai docenti italiani di diversi ordini e gradi scolastici. L'obiettivo metodo-

logico è duplice: da un lato, comprendere in profondità il sistema di convinzioni pedagogiche relative al feedback; dall'altro, analizzare le pratiche valutative effettivamente adottate, distinguendo fra feedback esplicito e implicito.

3. Risultati

Si presentano i primi risultati derivati dalle statistiche descrittive e affidabilità delle scale PAR e HAB

Scala	Sottoscala	Media	DS	α di Cronbach
PAR	Valore pedagogico del feedback	3,8	0,3	.82
PAR	Riconoscimento istituzionale	3,4	0,6	.75
HAB	Feedback esplicito	3,4	0,5	.78
HAB	Feedback implicito	2,4	0,8	.71

Tab. 1 – Riassunto dei risultati delle statistiche descrittive

Le risposte sono state raccolte su una scala Likert a quattro punti. Tutti i coefficienti di Cronbach superano la soglia di .70, indicando una buona affidabilità interna delle sottoscale.

Le analisi descrittive mostrano valori elevati per la dimensione PAR valore, che esprime il riconoscimento del feedback come leva fondamentale dell'apprendimento. Questo dato indica una solida cultura pedagogica condivisa: i docenti conoscono il valore formativo del feedback, lo riconoscono nella teoria e lo percepiscono come parte essenziale della loro professionalità.

La seconda dimensione, PAR istituzionale, presenta valori più moderati e variabili. Ciò suggerisce che il sistema scuola non sempre sostiene in modo coerente la cultura del feedback. I docenti credono profondamente nel feedback, ma tale convinzione non è necessariamente rafforzata dalle pratiche e dalle priorità dell'istituzione scolastica.

La scala HAB rivela due fattori. Il feedback esplicito include commenti scritti, correzioni, rubriche, restituzioni individuali. È il tipo di feedback che i docenti conoscono e utilizzano con maggiore sicurezza, coerente con le indicazioni ministeriali e la tradizione valutativa italiana. Il feedback implicito – segnali non verbali, modulazione delle consegne, orientamenti taciti, scaffolding relazionale – risulta invece la dimensione più fragile. La forte variabilità indica che alcuni

docenti padroneggiano forme di regolazione altamente sofisticate, mentre altri non le riconoscono nemmeno come feedback.

Sapere che il feedback è importante non significa saperlo realizzare in modo sofisticato. La distanza tra convinzioni e pratiche è un tema ben documentato anche nella ricerca sulle credenze degli insegnanti (Pajares, 1992; Fives & Buehl, 2012). Il risultato suggerisce che la traduzione delle idee in azioni richiede competenze più complesse della semplice consapevolezza teorica.

L'analisi di cluster ha permesso di individuare tre profili professionali distinti, che rappresentano altrettanti ecosistemi pedagogici nei quali convinzioni, pratiche e percezione del contesto scolastico si intrecciano in modo peculiare. Questa differenziazione conferma quanto sostenuto dalla letteratura sulle professionalità docenti: lungi dall'essere un gruppo omogeneo, i docenti tendono a sviluppare identità pedagogiche plurali, influenzate da fattori personali, organizzativi e culturali (Day & Gu, 2010; Fives & Buehl, 2012).

Pedagogicamente forti ma tecnologicamente cauti

I docenti hanno convinzioni pedagogiche solide riguardo al feedback: riconoscono la centralità della valutazione formativa, ne comprendono le funzioni regolative e dichiarano di utilizzarla in modo ricco sia nelle dimensioni esplicite sia in quelle implicite.

Tuttavia, mostrano un atteggiamento cauto verso le innovazioni non completamente consolidate, incluse le tecnologie digitali per il supporto al feedback. Tale cautela deve essere interpretata come espressione di un senso di responsabilità professionale. Tale visione è coerente con la prospettiva di Biesta (2010), che sottolinea la necessità di preservare il "giudizio professionale" di fronte a pressioni tecnologiche o performative.

Pratici moderati con scarso supporto istituzionale

Il secondo cluster è caratterizzato da livelli intermedi in tutte le dimensioni analizzate: convinzioni pedagogiche meno strutturate, uso moderato del feedback esplicito e un utilizzo particolarmente debole del feedback implicito. Ciò che distingue profondamente questo gruppo dagli altri è la bassa percezione del supporto istituzionale: questi docenti non sentono che la scuola valorizzi il feedback né che investa in una cultura condivisa della valutazione formativa.

La letteratura sull'efficacia scolastica indica che il clima istituzionale è un fattore determinante per sostenere l'engagement dei docenti e la qualità della didattica (Fullan, 2016).

Si tratta quindi di un profilo professionale potenzialmente vulnerabile, non

per mancanza di competenze, ma per assenza di un contesto che favorisca crescita e sperimentazione.

Innovatori valoriali

Il terzo profilo si distingue per convinzioni molto alte sul valore pedagogico del feedback, una forte percezione del supporto istituzionale e una spiccata apertura verso pratiche innovative, inclusa l'integrazione di tecnologie emergenti. A differenza del primo cluster, questi docenti interpretano l'innovazione tecnologica non come un rischio per la relazione educativa, ma come una *opportunità* per ampliare il potenziale trasformativo del feedback.

Il loro orientamento è coerente con le ricerche sulla *teacher innovation mindset* (Zhao, 2012), secondo cui l'apertura al nuovo deriva da una combinazione di fiducia pedagogica, sostegno organizzativo e disponibilità a esplorare forme di didattica potenziata. Questo gruppo appare quindi particolarmente predisposto a sperimentare strumenti come l'intelligenza artificiale, purché in coerenza con valori educativi chiari.

Nel loro insieme, i tre profili delineano una mappa ricca e articolata della professione docente contemporanea. Le combinazioni tra convinzioni, pratiche e percezione del contesto rivelano una forte cultura pedagogica che non implica automaticamente apertura tecnologica; l'assenza di un clima istituzionale favorevole indebolisce sia le convinzioni sia le pratiche.L'innovazione si sviluppa dall'interazione tra valori professionali profondi e sostegno organizzativo; e il feedback implicito si conferma un "territorio formativo critico", presente solo nei profili più maturi.

Queste evidenze suggeriscono l'urgenza di percorsi formativi differenziati, costruiti non su un modello unico, ma calibrati sui bisogni dei diversi ecosistemi professionali. La professionalità docente non evolve in modo uniforme: occorre quindi riconoscerne le sfumature per sostenere processi di sviluppo autentici e duraturi.

L'EFA ha individuato due fattori chiari:

Item (contenuto sintetico)	Valore pedagogico	Riconoscimento istituzionale
Il feedback favorisce la comprensione degli errori	.72	—
Il feedback guida il miglioramento dell'apprendimento	.68	—
Il feedback ha una funzione regolativa nel processo didattico	.75	—

Item (contenuto sintetico)	Valore pedagogico	Riconoscimento istituzionale
La scuola valorizza il feedback nella valutazione	—	.70
Colleghi e leadership sostengono l'uso del feedback	—	.66

Tab. 2 – Carichi fattoriali delle scale PAR

Sono riportati solo i carichi fattoriali I .40. Metodo di estrazione: Principal Axis Factoring; rotazione: oblimin.

La Tabella 2 riporta i carichi fattoriali emersi dall'analisi fattoriale esplorativa delle scale PAR, per cui si individuano due fattori distinti e chiaramente interpretabili, coerenti con il modello teorico: il valore pedagogico del feedback e il riconoscimento istituzionale del feedback. Gli item presentano carichi elevati sul fattore di riferimento e carichi trascurabili sull'altro fattore, indicando una buona validità discriminante.

Analogamente, per la scala HAB, l'analisi ha confermato la distinzione tra pratiche di feedback esplicito e implicito. I carichi fattoriali risultano superiori alla soglia di .40 e coerenti con la distinzione concettuale proposta dalla letteratura internazionale. Nel complesso, la struttura fattoriale conferma l'articolazione del feedback lungo un asse valoriale, che orienta le convinzioni dei docenti, e un asse pratico, che regola il comportamento valutativo effettivamente adottato.

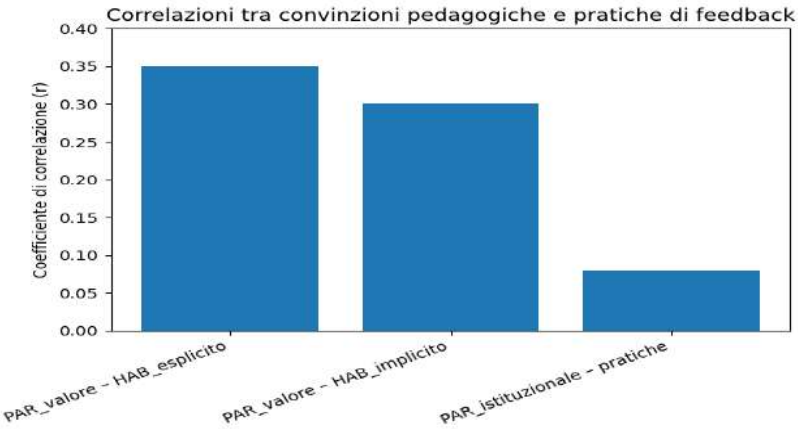
Item (contenuto sintetico)	Feedback esplicito	Feedback implicito
Fornisco commenti scritti sui lavori degli studenti	.74	—
Utilizzo rubriche o criteri espliciti di valutazione	.69	—
Fornisco indicazioni di miglioramento (feedforward)	.71	—
Regolo le richieste in base alle risposte degli studenti	—	.73
Uso segnali non verbali per orientare l'apprendimento	—	.68
Offro supporto relazionale durante l'attività	—	.65

Tab. 3 – Carichi fattoriali della scala HAB – Pratiche valutative ricorrenti. Sono riportati solo i carichi fattoriali I .40. Metodo di estrazione: Principal Axis Factoring; rotazione: oblimin.

Asse	Fattore	Tipologia di feedback	Caratteristiche principali
Asse pratico	Fattore 1	Feedback esplicito	Pratiche dichiarate e osservabili; commenti scritti, correzioni, rubriche, feedforward
Asse pratico	Fattore 2	Feedback implicito	Strategie tacite, relazionali e situate; segnali non verbali, modulazione delle richieste, scaffolding relazionale

Tab. 4 – Struttura fattoriale delle pratiche di feedback (EFA)

Le analisi di correlazione mostrano una relazione moderata tra il valore pedagogico attribuito al feedback e le pratiche di feedback esplicito (r T .35), mentre la relazione con le pratiche di feedback implicito risulta più debole (r T .30). Le correlazioni tra il riconoscimento istituzionale del feedback e le pratiche valutative risultano invece molto basse.



Graf. 1 – Correlazioni tra convinzioni pedagogiche e pratiche di feedback

Nel loro insieme, questi risultati indicano che, sebbene i docenti condividano convinzioni pedagogiche solide sul valore del feedback, tali convinzioni non si traducono automaticamente in pratiche valutative complesse e sofisticate. Questo risultato fornisce una conferma empirica del *theory–practice gap*, secondo cui gli insegnanti tendono a "sapere più di quanto riescano ad agire consapevolmente" (Loughran, 2010).

L'analisi complessiva dei dati restituisce un quadro articolato della professionalità docente, mettendo in luce dinamiche complesse tra convinzioni pedago-

giche, pratiche valutative e contesto istituzionale. I risultati non possono essere letti in modo lineare, ma richiedono un'interpretazione sistemica, capace di cogliere interazioni, tensioni e asimmetrie tra i diversi livelli analizzati.

Tra tutte le dimensioni analizzate, il feedback implicito emerge come l'area più fragile e discriminante. I punteggi medi bassi e l'elevata varianza indicano che questa forma di feedback è praticata in modo discontinuo e profondamente diseguale tra i docenti. Mentre alcuni insegnanti dimostrano di padroneggiare modalità di regolazione sofisticate – basate sull'osservazione, sull'adattamento in tempo reale e sulla modulazione situata delle interazioni – altri non riconoscono tali pratiche come elementi intenzionali del proprio repertorio professionale.

Questa difficoltà può essere interpretata alla luce della natura stessa del feedback implicito, che la letteratura definisce come tacito, relazionale e incorporato nell'azione (Carless & Boud, 2018). A differenza del feedback esplicito, esso non è facilmente codificabile né formalizzabile e richiede al docente un alto livello di consapevolezza metacognitiva e di sensibilità pedagogica. Il fatto che questa dimensione rappresenti il principale punto di frattura tra i profili professionali suggerisce che lo sviluppo del feedback implicito costituisce una delle sfide più rilevanti per la formazione docente contemporanea.

Un risultato particolarmente rilevante riguarda il ruolo del contesto istituzionale. La variabile PAR istituzionale si dimostra centrale nel differenziare i profili di docenti e nel sostenere, o al contrario indebolire, l'allineamento tra convinzioni e pratiche. I docenti che percepiscono un forte riconoscimento istituzionale del feedback tendono a mostrare maggiore apertura alla sperimentazione.

Questo dato conferma l'importanza del clima organizzativo e della leadership scolastica nei processi di cambiamento educativo (Fullan, 2016; Leithwood et al., 2020). In assenza di un contesto che valorizzi, legittimi e accompagni le pratiche formative, anche docenti con convinzioni solide possono faticare a mantenere strategie complesse e riflessive nel lungo periodo.

La professione docente è certamente un ecosistema complesso.

Nel loro insieme, i risultati suggeriscono che la professione docente non può essere interpretata come un blocco omogeneo. Le differenze osservate tra i profili non sono riconducibili a variabili demografiche tradizionali, come età anagrafica, anni di servizio o ordine di scuola, ma riflettono differenze culturali, valoriali e operative profondamente radicate nelle traiettorie professionali individuali e nei contesti organizzativi di riferimento.

Questa evidenza rafforza una visione ecologica della professionalità docente (Day & Gu, 2010; Hargreaves & Fullan, 2012), secondo cui sviluppo professionale, identità e pratiche emergono dall'interazione dinamica tra individuo e

sistema. I docenti non evolvono in modo lineare né uniforme: essi si muovono all'interno di ecosistemi professionali differenti, che richiedono strategie di supporto e formazione diversificate.

4. Conclusione

I risultati rivelano un quadro articolato: i docenti credono fortemente nel feedback, ma tale convinzione non sempre si traduce in pratiche complesse, soprattutto nella dimensione implicita. È proprio qui che emergono margini di sviluppo professionale.

L'interpretazione integrata dei risultati mostra che la qualità del feedback nella scuola non dipende esclusivamente dalle competenze individuali dei docenti, ma da un equilibrio delicato tra convinzioni pedagogiche, consapevolezza pratica e condizioni istituzionali. La sfida principale non consiste nel "convincere" i docenti dell'importanza del feedback – poiché tale convinzione è già ampiamente diffusa – bensì nel sostenere la trasformazione delle convinzioni in pratiche raffinate, soprattutto nelle loro forme implicite e relazionali, attraverso ambienti scolastici che riconoscano e valorizzino la complessità del lavoro docente.

Il punto critico non è quindi l'assenza di valori pedagogici, ma la difficoltà di trasformarli in comportamenti didattici coerenti. Questo divario è amplificato dalla percezione, non sempre positiva, del sostegno istituzionale, che risulta fattore decisivo per alimentare pratiche di qualità e continuità.

Infine, la differenziazione in cluster suggerisce che non tutti i docenti hanno bisogno delle stesse cose: alcuni richiedono formazione tecnica sulle pratiche implicite, altri necessitano di un rafforzamento valoriale, altri ancora di un ambiente scolastico più favorevole alla sperimentazione. Questi risultati indicano la necessità di percorsi formativi mirati e differenziati, centrati sulla feedback literacy, e di un più deciso investimento istituzionale nella cultura della valutazione formativa.

Riferimenti bibliografici

Biesta, G. (2010). *Good education in an age of measurement.* Routledge.
Brookhart, S. (2008). *How to give effective feedback to your students.* ASCD.
Carless, D., & Boud, D. (2018). The development of student feedback literacy. *Assessment & Evaluation in Higher Education.*

Carless, D., & Winstone, N. (2023). *Teacher feedback: A relational approach.* Routledge.
Fives, H., & Buehl, M. (2012). Teachers' beliefs and their importance. In *APA Educational Psychology Handbook*.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.
Laurillard, D. (2012). *Teaching as a design science.* Routledge.
Pajares, M. (1992). Teachers' beliefs and educational practice. *Review of Educational Research.*
Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.
Wiliam, D. (2011). *Embedded formative assessment.* Solution Tree Press.

II.27
Dall'autovalutazione al feedback personalizzato: l'integrazione dell'intelligenza artificiale nella costruzione del bilancio di competenze nei percorsi di abilitazione dei docenti

Sofia Di Crisci
Docente in utilizzo come Tutor Coordinatore presso Università di Trento
sofia.dicrisci@unitn.it
Cristiana Bianchi
Docente in utilizzo come Tutor Coordinatore presso Università di Trento
cristiana.bianchi@unitn.it

Abstract

Il contributo documenta un'esperienza di integrazione dell'IA nei processi di valutazione formativa condotta nell'ambito dei Percorsi Abilitanti per docenti della scuola secondaria presso l'Università di Trento. Il percorso ha coinvolto quaranta corsisti nella costruzione collaborativa di un bilancio di competenze attraverso l'utilizzo di strumenti di intelligenza artificiale generativa quali NotebookLM, Claude.ai e Napkin.ai. L'esperienza ha consentito di esplorare le potenzialità di un utilizzo pedagogicamente fondato dell'intelligenza artificiale nella personalizzazione dei feedback formativi, nell'analisi sistematica delle competenze professionali emergenti e nella progettazione responsiva dei percorsi di tirocinio. Il percorso realizzato evidenzia come un'integrazione critica e consapevole delle tecnologie emergenti possa arricchire significativamente i dispositivi valutativi favorendo riflessività, metacognizione e personalizzazione in tempi compatibili con quelli dei percorsi formativi, a condizione di mantenere centrale la responsabilità pedagogica del formatore.

Parole chiave: valutazione formativa; intelligenza artificiale; bilancio di competenze; formazione insegnanti; feedback personalizzato.

This paper documents an experience of integrating AI into formative assessment processes within teacher qualification pathways for secondary school teachers at the University of Trento. The pathway involved forty trainees in the collaborative construction of a competency portfolio through the use of generative artificial intelligence tools such as NotebookLM, Claude.ai, and Napkin.ai. The experience enabled exploration of the potential of pedagogically informed use of artificial intelligence in personalizing formative feedback, systematically analyzing

emerging professional competencies, and designing adaptive practicum experiences. The implemented pathway demonstrates how critical and intentional integration of emerging technologies can substantially enrich assessment practices by promoting reflexivity, metacognition, and personalization within timeframes compatible with teacher education programs, provided that the educator's pedagogical responsibility remains central.

Keywords: formative assessment; artificial intelligence; competency portfolio; teacher education; personalized feedback.

Introduzione

Il mondo dell'educazione sta attraversando una fase di trasformazione che ridefinisce le pratiche di formazione dei futuri docenti. L'IA generativa offre nuove possibilità per ripensare la valutazione formativa e l'autovalutazione, aprendo scenari inediti per l'analisi dei processi di apprendimento e la personalizzazione dei percorsi formativi[1].

Il presente contributo documenta un'esperienza sperimentale condotta da due tutor coordinatrici presso l'Università di Trento nell'ambito dei Percorsi Abilitanti per docenti della scuola secondaria, un percorso incentrato sulla costruzione collaborativa del bilancio di competenze e realizzato mediante l'integrazione di strumenti di IA. L'esperienza muove dalla convinzione che la valutazione delle competenze professionali dei docenti in formazione richieda approcci capaci di intrecciare riflessione e metacognizione[2], favorendo quella personalizzazione che con i metodi tradizionali resta spesso difficilmente raggiungibile. In qualità di tutor coordinatrici, l'iniziativa è stata promossa in forma sperimentale, con particolare attenzione alla costruzione di dispositivi valutativi orientati alla crescita professionale dei corsisti. Un elemento distintivo dell'intero percorso è stata l'attenzione sistematica alla trasferibilità delle pratiche proposte. Ogni attività è stata progettata con duplice finalità: sostenere lo sviluppo professionale dei corsisti e proporre modelli operativi adattabili nei contesti scolastici. La consapevolezza che i corsisti stessero sperimentando in prima persona pratiche valutative ha rappresentato un'opportunità preziosa per riflettere criti-

1 UNESCO (2021). AI and education: Guidance for policy-makers. Paris: UNESCO Publishing.

2 Schön, D. A. (1983). The reflective practitioner: How professionals think in action. New York: Basic Books.

camente sul ruolo della valutazione nell'apprendimento[3]. L'integrazione di strumenti di IA è derivata dalla volontà di esplorarne le potenzialità nel sostenere la riflessione metacognitiva e personalizzare i feedback, mantenendo al centro la dimensione pedagogica e la responsabilità professionale del formatore. Sono stati utilizzati strumenti di IA generativa per l'analisi documentale, l'elaborazione di feedback individualizzati e la creazione di visualizzazioni sintetiche.

Quadro teorico di riferimento

Il presente lavoro si colloca nella tradizione di studi sulla valutazione per competenze e sulla formazione docenti. La valutazione autentica delle competenze richiede di rendere visibili prestazioni complesse, favorendo processi di autovalutazione e consapevolezza metacognitiva (Castoldi, 2016)[4]. Questo approccio privilegia dimensioni qualitative e processuali, orientando verso forme di valutazione che sostengano effettivamente la crescita professionale e superino la logica puramente quantitativa (Corsini, 2023), trovando fondamento negli strumenti dell'evidence-based education (Trinchero, 2023). La competenza professionale docente viene concettualizzata come mobilitazione integrata di risorse cognitive, affettive e sociali in contesti situati (Pellerey, 2004)[5]. Questa prospettiva complessa e multidimensionale ha guidato la costruzione collaborativa del framework di riferimento, privilegiando una visione della professionalità che supera l'idea di semplice possesso di conoscenze disciplinari. Il costrutto di riflessione, elemento centrale del percorso, si richiama alla tradizione del professionista riflessivo[6] e alle teorie dell'apprendimento trasformativo (Mezirow, 2003), secondo cui la riflessione critica sulla pratica costituisce il fondamento per apprendimenti autentici e trasformazioni profonde.

Gli studi sul feedback formativo hanno orientato in modo decisivo la progettazione delle restituzioni valutative. Un feedback efficace deve rispondere a tre domande chiave, dove sto andando, come sto procedendo, quali sono i passi successivi (Hattie & Timperley, 2007).[7] L'efficacia dipende dalla capacità di fa-

3 Wiliam, D. (2011). What is assessment for learning? Studies in Educational Evaluation, 37(1), 3-14.

4 La valutazione autentica si focalizza su compiti reali e contestualizzati che richiedono l'applicazione di conoscenze e competenze in situazioni complesse.

5 Questa concezione della competenza supera l'approccio comportamentista per abbracciare una visione olistica che considera la persona nella sua interezza.

6 Schön, D. A. (1983), op. cit.

7 Queste tre domande strutturano il framework concettuale del feedback formativo efficace, orientando la riflessione sia del formatore che del discente.

cilitare l'autoregolazione e la metacognizione (Nicol & Macfarlane-Dick, 2006), fornendo indicazioni specifiche che identifichino passi raggiungibili piuttosto che traguardi che appaiano irraggiungibili. Particolarmente rilevante è il potere del feedback interno e dei processi comparativi naturali (Nicol, 2021)[8]. La dimensione emotiva del feedback è cruciale poiché influenza significativamente l'apprendimento (Lipnevich et al., 2021),[9] rendendo necessario calibrare il tono e la struttura considerando non solo la dimensione cognitiva ma anche quella affettiva e motivazionale.

L'integrazione dell'IA nell'educazione richiede sviluppo di una riflessione critica sulle dimensioni etiche e metodologiche.[10] L'intelligenza artificiale deve supportare, non sostituire, il giudizio professionale del formatore (Holmes et al., 2019; Popenici & Kerr, 2017). L'uso pedagogicamente fondato delle tecnologie richiede consapevolezza metodologica e capacità di progettazione didattica (DigCompEdu)[11]. L'uso consapevole dei dati richiede competenze specifiche di data literacy (Raffaghelli, 2019) e ripensamento dei processi valutativi (Bearman et al., 2020).

Contesto e partecipanti

L'esperienza si è svolta nei percorsi abilitanti istituiti con il decreto ministeriale n. 108 del 2023[12]. La sperimentazione ha coinvolto quaranta corsisti, venti della classe A022 (italiano, storia, geografia) e venti della classe A028 (matematica e scienze). I partecipanti presentavano profili diversificati per esperienze pregresse, percorsi formativi e familiarità con gli strumenti digitali. Il contesto si è configurato come laboratorio per sperimentare varie modalità di valutazione, considerando la duplice posizione dei corsisti quali discenti e futuri professionisti.

8 Il concetto di "feedback interno" evidenzia come gli apprendenti generino autonomamente giudizi valutativi attraverso processi di confronto cognitivo.

9 La ricerca dimostra che feedback percepiti come eccessivamente critici o giudicanti possono generare emozioni negative che inibiscono l'apprendimento.

10 Holmes, W., Bialik, M., & Fadel, C. (2019) sottolineano come l'AI debba essere progettata e utilizzata per amplificare le capacità umane, non per sostituirle.

11 Redecker, C., & Punie, Y. (2017). European framework for the digital competence of educators: DigCompEdu. Luxembourg: Publications Office of the European Union.

12 Ministero dell';Istruzione e del Merito (2023). Decreto Ministeriale n. 108 del 4 agosto 2023.

Descrizione del percorso formativo

Il percorso ha preso avvio dalla condivisione con i corsisti di un corpus documentale sulla professionalità docente comprendente: gli Australian Professional Standards for Teachers[13], il Contratto Collettivo Nazionale, documenti sullo sviluppo professionale continuo, il dossier MIUR e l'allegato A del decreto ministeriale. È stato introdotto NotebookLM per l'analisi dei documenti, sviluppando competenze di ricerca e elaborazione critica.

Successivamente i corsisti hanno operato in gruppi per elaborare un quadro di riferimento delle competenze professionali. I gruppi hanno identificato dimensioni, indicatori e livelli di padronanza, innescando processi di negoziazione e confronto[14].

Le produzioni dei gruppi sono state sintetizzate mediante Claude.ai, che ha consentito di identificare convergenze, integrare prospettive e costruire un framework condiviso con cinque dimensioni della professionalità docente, competenze disciplinari e didattiche, competenze relazionali e comunicative, competenze valutative, competenze organizzative e gestionali, sviluppo professionale continuo. Per ciascuna competenza sono stati identificati indicatori specifici e livelli progressivi di padronanza. Ciascun corsista ha poi compilato il proprio bilancio di competenze iniziale, avviando un processo di autovalutazione riflessiva[15]. Lo schema di bilancio proposto richiede di posizionarsi rispetto a ciascuna dimensione del framework condiviso, argomentando il proprio livello attraverso evidenze concrete tratte dall'esperienza formativa e professionale pregressa. Questa fase ha richiesto investimento riflessivo significativo, sollecitando metacognizione e consapevolezza delle proprie risorse e aree di sviluppo, costituendo un momento cruciale per identificare obiettivi di crescita da perseguire attraverso i tirocini e la formazione teorica.

I quaranta bilanci di competenza sono stati analizzati sistematicamente mediante NotebookLM. Tale operazione ha consentito di identificare pattern ricorrenti, aree di sviluppo comuni, specificità individuali e differenze tra le due classi di concorso. È emerso che i corsisti manifestavano, in entrambi i gruppi, maggiore sicurezza nelle competenze disciplinari rispetto a quelle didattiche,

13 Teacher Education Ministerial Advisory Group (2014). Australian Professional Standards for Teachers. Melbourne: Australian Institute for Teaching and School Leadership.

14 La co-costruzione di criteri valutativi è una pratica sostenuta dalla letteratura sull'assessment as learning. Cfr. Grion et al. (2019).

15 Il bilancio di competenze come strumento riflessivo ha radici nella tradizione francofona della formation professionnelle e nella pratica del portfolio nell'educazione degli adulti.

metodologiche e valutative. Per l'elaborazione dei feedback personalizzati sono stati integrati NotebookLM e Claude.ai, ottenendo calibrazione precisa e reale personalizzazione. La progettazione dei prompt per la composizione dei feedback è stata orientata da scelte pedagogiche precise, valorizzare i punti di forza emersi dal bilancio, descrivere con precisione le criticità in relazione al framework condiviso, fornire suggerimenti pratici che aiutassero a identificare piccoli passi possibili verso il miglioramento piuttosto che traguardi irraggiungibili[16].

Un esempio del prompt utilizzato per l'elaborazione dei feedback attraverso Claude.ai può chiarire l'approccio adottato. Il prompt strutturato conteneva le seguenti indicazioni: "Agisci come tutor coordinatore esperto nella formazione iniziale degli insegnanti, con competenze in mentoring educativo e sviluppo professionale docente. Il tuo obiettivo è fornire un feedback personalizzato, formativo e incoraggiante basato sul bilancio delle competenze che trovi allegato scritto dal corsista, utilizzando un approccio dialogico e orientato alla crescita professionale. Approccio generale, adotta un tono empatico, professionale ma caloroso, orientato a costruire fiducia e senso di autoefficacia. Evita giudizi valutativi e privilegia un linguaggio descrittivo e orientato al miglioramento continuo. Riferisciti, quando pertinente, ai framework europei delle competenze chiave per l'apprendimento permanente (2018)[17] e al framework DigCompEdu per le competenze digitali degli educatori. Struttura del feedback, inizia valorizzando i traguardi già raggiunti, riconoscendo con specificità le competenze consolidate e collegandole a pratiche didattiche efficaci. Valorizza il lavoro riflessivo compiuto dal corsista nell'individuare le aree in crescita, sottolineando l'importanza della metacognizione nella professione docente. Trasforma le aree da acquisire in opportunità di sviluppo professionale, presentandole come sfide stimolanti e tappe naturali del percorso formativo, non come mancanze o deficit. Per ciascuna area da consolidare o acquisire, proponi almeno due-tre strategie concrete e differenziate, per il miglioramento e suggerisci letture professionali mirate con indicazione di testi di riferimento in ambito pedagogico e didattico. Chiudi con una frase motivante che sottolinei il valore della formazione continua come dimensione costitutiva dell'identità professionale docente. Incoraggia il corsista a considerare il bilancio delle competenze come uno strumento dinamico, da rivisitare periodicamente per monitorare la propria crescita professionale. Ribadisci la disponibilità a supportare il percorso formativo con ulteriori

16 Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153-189, evidenzia l'importanza di fornire indicazioni specifiche e attuabili.

17 Consiglio dell'Unione Europea (2018). Raccomandazione del Consiglio del 22 maggio 2018 relativa alle competenze chiave per l'apprendimento permanente.

confronti e momenti di supervisione riflessiva";. Per ciascun corsista è stato elaborato un feedback rispettoso di questa struttura. L'IA ha consentito una standardizzazione degli esiti per quanto riguarda la forma e una personalizzazione significativa, tuttavia tutti i feedback sono stati rivisti, validati, integrati e talvolta significativamente modificati dalle tutor, mantenendo costantemente centrale la responsabilità pedagogica.

Per garantire la protezione dei dati personali e il rispetto della privacy,[18] tutti i bilanci di competenza sono stati preventivamente anonimizzati prima del caricamento su Claude.ai. L'elaborazione dei feedback è stata condotta individualmente, un bilancio alla volta, evitando l'inserimento simultaneo di materiali riferiti a più corsisti. Questa procedura ha consentito di minimizzare i rischi connessi al trattamento dei dati attraverso piattaforme di IA, assicurando che nessuna informazione identificativa venisse elaborata o conservata negli strumenti tecnologici utilizzati.

Mediante Napkin.ai sono state costruite sintesi visive per ciascuna classe, che hanno permesso di visualizzare graficamente i profili emergenti e le traiettorie di sviluppo.[19] Le infografiche hanno restituito al gruppo una visione d'insieme dei punti di forza collettivi e delle aree su cui concentrare l'attenzione. Il gruppo A022 ha manifestato forte interesse per la didattica collaborativa e l'analisi testuale, mentre il gruppo A028 ha evidenziato priorità verso metodologie laboratoriali e interdisciplinarità. La sintesi degli esiti generali si è rivelata strumento prezioso per i tutor nella progettazione del tirocinio indiretto e per i corsisti nella stesura del progetto di tirocinio diretto, consentendo di individuare con precisione i contenuti e le metodologie delle attività laboratoriali, focalizzare l'attenzione sulle dimensioni professionali emerse come maggiormente critiche e orientare i corsisti nella scelta delle esperienze di tirocinio diretto.

Indagine qualitativa sulla percezione del percorso

Al termine del percorso di tirocinio indiretto è stata condotta un'indagine qualitativa per rilevare la percezione dei corsisti relativamente all'efficacia del percorso formativo, con particolare attenzione ai feedback personalizzati ricevuti e

18 Regolamento UE 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 (GDPR).

19 Le visualizzazioni dei dati supportano processi di data literacy e promuovono interpretazioni condivise dei risultati valutativi.

all'integrazione degli strumenti di IA[20]. L'indagine è stata realizzata attraverso un questionario anonimo somministrato mediante form online, articolato in cinque sezioni: gradimento generale, efficacia dei temi trattati, efficacia dei feedback, confronto con il tutor coordinatore, osservazioni e suggerimenti. Il questionario ha utilizzato scale Likert a quattro punti (configurandosi come una scala a "scelta forzata" eliminando quindi l'opzione neutra o intermedia) per le domande chiuse e ha previsto uno spazio per risposte aperte nella sezione finale.

Le domande pertinenti con il tema del presente contributo hanno riguardato specificatamente l'efficacia dei feedback personalizzati elaborati con il supporto dell'intelligenza artificiale. Hanno risposto 34 corsisti su 40 (85% di tasso di risposta). In particolare è stato chiesto ai corsisti di valutare se i feedback ricevuti dalla tutor coordinatrice fossero stati chiari e comprensibili, se avessero contribuito a migliorare la riflessione sulle pratiche didattiche, se le osservazioni e i suggerimenti ricevuti fossero stati costruttivi e orientati al miglioramento, e se i feedback avessero stimolato una maggiore consapevolezza delle proprie competenze e aree di miglioramento.

I risultati dell'indagine evidenziano che l'85,3% dei corsisti ha valutato i feedback ricevuti come chiari o molto chiari, mentre il 79,4% ha ritenuto che i feedback abbiano contribuito molto o moltissimo al miglioramento della riflessione sulle pratiche didattiche. Per quanto riguarda la costruttività delle osservazioni e dei suggerimenti ricevuti, il 97,1% dei corsisti li ha valutati come costruttivi o molto costruttivi. Particolarmente significativo è il dato relativo alla consapevolezza metacognitiva: il 94,1% dei corsisti ha dichiarato che i feedback hanno stimolato molto o moltissimo una maggiore consapevolezza delle proprie competenze e aree di miglioramento. Infine, il 73,5% dei partecipanti ha valutato il confronto con la tutor coordinatrice come un'esperienza arricchente o molto arricchente.

Dalle risposte aperte emergono alcune considerazioni qualitative particolarmente rilevanti. Un corsista ha scritto: "Ciò che ho apprezzato di più è la passione, la disponibilità e la dedizione mostrata dalla tutor. [...] Il suo reale desiderio di portarci al miglioramento e la cura che ha messo nel farlo sono tangibili". Un altro ha commentato: "Ho apprezzato la sua grande professionalità, il supporto costante e le lezioni sempre chiare, coinvolgenti e ricche di spunti. Grazie a lei ho imparato davvero tanto!" Inoltre, è emerso che il confronto con la tutor su situazioni reali "mi ha dato strategie aggiuntive per la gestione reale

20 L'utilizzo di scale a scelta forzata evita il "central tendency bias" tipico delle scale con punto medio, obbligando i rispondenti a prendere posizione.

di contesti difficili" e che "i feedback mi hanno aiutato a capire come interagire in modo più efficace nella valutazione dei miei studenti". Questi elementi confermano l'efficacia dell'approccio adottato nella progettazione dei feedback attraverso l'intelligenza artificiale, evidenziando come la combinazione tra personalizzazione tecnologica e supervisione pedagogica delle tutor abbia generato restituzioni valutative percepite come autenticamente formative e orientate alla crescita professionale.

Analisi e riflessioni

L'utilizzo di NotebookLM ha favorito lo sviluppo di competenze di ricerca e sintesi[21]. L' analisi sistematica dei bilanci ha fatto emergere pattern difficilmente rilevabili con analisi manuale. La personalizzazione dei feedback è risultata efficace. Le visualizzazioni degli esiti accorpati hanno favorito la consapevolezza collettiva. L'attenzione alla trasferibilità ha favorito duplice consapevolezza nei corsisti. Rimangono alcune criticità come il rischio di delega acritica alla tecnologia, infatti l'utilizzo dell'IA richiede che il formatore mantenga il controllo pedagogico[22]. Una seconda criticità riguarda la trasparenza, è fondamentale che i corsisti siano consapevoli delle modalità di utilizzo dell'intelligenza artificiale. Una terza questione riguarda la protezione dei dati e la privacy, affrontata attraverso l'anonimizzazione preventiva dei materiali e l'elaborazione individuale dei feedback. Infine, è necessaria riflessione critica sulla validità pedagogica dei feedback elaborati con supporto dell'IA.

Conclusioni

L'esperienza documentata ha consentito di esplorare modalità integrate con l'IA nei processi di valutazione formativa e autovalutazione nella formazione iniziale dei docenti. Si è osservato che un utilizzo consapevole, critico e pedagogicamente fondato delle tecnologie emergenti può arricchire significativamente i dispositivi valutativi, favorendo personalizzazione, riflessività e metacognizione in tempi

21 Gli strumenti di IA generativa come NotebookLM consentono analisi documentali che superano le capacità di sintesi umana per velocità e sistematicità, pur richiedendo sempre interpretazione pedagogica.

22 Holmes et al. (2019) sottolineano come l'AI debba rimanere uno strumento al servizio del giudizio professionale dell'educatore, mai un sostituto.

compatibili con quelli del percorso formativo. I feedback personalizzati sono stati percepiti dai corsisti come particolarmente utili, e hanno favorito riflessioni approfondite sulle proprie competenze orientando efficacemente le scelte relative al tirocinio. La progettazione accurata dei prompt si è rivelata cruciale per garantire la qualità pedagogica.[23]

La restituzione degli esiti ha orientato la progettazione del tirocinio e ha promosso nei corsisti una prospettiva meta-professionale, stimolando interesse verso l'uso pedagogicamente fondato degli strumenti di IA.

L'esperienza evidenzia la necessità di mantenere centrale la responsabilità pedagogica del formatore. Validazione critica, integrazione con la conoscenza esperta e trasparenza costituiscono condizioni imprescindibili. L'integrazione dell'IA nella valutazione formativa rappresenta una prospettiva promettente, a condizione che sia sostenuta da solide basi pedagogiche e attenzione alle dimensioni etiche. L'IA può essere un alleato prezioso nella personalizzazione, ma solo se rimane al servizio di una visione dell'educazione che pone al centro la persona e la costruzione condivisa di significati.

Riferimenti bibliografici

Bearman, M., Dawson, P., Boud, D., Bennett, S., Hall, M., & Molloy, E. (2020). Support for assessment practice: Developing the assessment design decisions framework. *Teaching in Higher Education*, 27(1), 19-36.

Castoldi, M. (2016). *Valutare e certificare le competenze.* Roma: Carocci.

Consiglio dell'Unione Europea (2018). *Raccomandazione del Consiglio del 22 maggio 2018 relativa alle competenze chiave per l'apprendimento permanente*. Gazzetta ufficiale dell'Unione europea, C 189.

Corsini, C. (2023). *La valutazione che educa. Liberare insegnamento e apprendimento dalla tirannia del voto*. Milano: FrancoAngeli.

Grion, V., Cook-Sather, A., Serbati, A., Motta, L., & Colussi, E. (2019). Towards a scholarship of comparison: Recognising the scholarship involved in comparative student-staff partnerships in assessment. *Student Engagement in Higher Education Journal,* 2(3), 201-215.

Grion, V., Serbati, A., Tino, C., & Nicol, D. (2021). Ripensare la teoria e le pratiche della valutazione in ottica inclusiva: Le potenzialità offerte dai processi di comparazione. *Italian Journal of Educational Research*, 14(26), 12-28.

23 La progettazione dei prompt rappresenta un aspetto cruciale dell'utilizzo pedagogico dell'IA, richiedendo competenze sia tecniche che didattiche. Cfr. Redecker & Punie (2017).

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Boston: Center for Curriculum Redesign.

Lipnevich, A. A., Murano, D., Krannich, M., & Goetz, T. (2021). Should I grade or should I comment? How to enhance students motivation and self-regulation through feedback. In L. Chen, S. Chih (Eds.), *International handbook of academic motivation and learning* (pp. 197-221). New York: Routledge.

Mezirow, J. (2003). Transformative learning as discourse. *Journal of Transformative Education*, 1(1), 58-63.

Ministero dell'Istruzione e del Merito (2023). Decreto Ministeriale n. 108 del 4 agosto 2023 - Percorsi universitari e accademici di formazione iniziale e continua e sistema di formazione abilitante all'insegnamento nella scuola secondaria.

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education*, 46(5), 756-778.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218.

Pellerey, M. (2004). *Le competenze individuali e il portfolio*. Scandicci: La Nuova Italia.

Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1), 1-13.

Raffaghelli, J. E. (2019). *Educators data literacy: Building a critical understanding of learning analytics*. London: Routledge.

Redecker, C., & Punie, Y. (2017). *European framework for the digital competence of educators: DigCompEdu*. Luxembourg: Publications Office of the European Union.

Regolamento UE 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali (GDPR).

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153-189.

Teacher Education Ministerial Advisory Group (2014). *Australian Professional Standards for Teachers*. Melbourne: Australian Institute for Teaching and School Leadership.

Trinchero, R. (2023). *Evidence Based Education: I dati a supporto della ricerca didattica e della pratica educativa.* Milano: Mondadori Università.

UNESCO (2021). *AI and education: Guidance for policy-makers.* Paris: UNESCO Publishing.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation,* 37(1), 3-14.

II.28
Dall'uso delle rubriche all'IA generativa: un chatbot per il feedback intermedio dei Project Work[1]

Elisabetta Lucia De Marco
Professoressa Associata, Università Pegaso, elisabettalucia.demarco@unipegaso.it
Marilena di Padova
Professoressa Associata, Università Pegaso, marilena.dipadova@unipegaso.it
Anna Dipace
Docente Ordinaria di Pedagogia Sperimentale, Università Pegaso, anna.dipace@unipegaso.it

Abstract

Il contributo presenta la progettazione di un chatbot conversazionale di intelligenza artificiale generativa *rubric-aware*, a supporto del feedback formativo docente durante l'elaborazione della prova finale (Project Work) degli studenti nei corsi di laurea triennali. L'esperienza si fonda sui principi della valutazione formativa e sull'uso delle rubriche come dispositivi di trasparenza, coerenza valutativa e mediazione dell'apprendimento. Il chatbot è stato progettato secondo un modello *human-in-the-loop* che preserva la responsabilità valutativa del docente. La progettazione include la codifica della rubrica ufficiale, la definizione docente dei prompt e un'architettura GenAI con *retrieval* di contenuti contestuali. I risultati evidenziano il potenziale dello strumento nel migliorare qualità, coerenza e tempestività del feedback intermedio, favorendo *feedback literacy* e apprendimento autoregolato.

Parole chiave: Intelligenza artificiale generativa; feedback formativo; chatbot basato su rubriche di valutazione; lavoro progettuale.

The paper presents the design of a rubric-aware generative AI conversational chatbot to support teacher-provided formative feedback during the development of the final project (Project Work) by students in undergraduate degree programmes. The experience is grounded in the principles of formative assessment and the use of rubrics as tools for transparency, evaluative coherence, and learning

1 Il contributo è il risultato di un lavoro condiviso. Ad ogni modo, ai fini dell'attribuzione Elisabetta Lucia De Marco ha scritto i paragrafi 1 e 3; Marilena Di Padova i paragrafi 4 e 5; Anna Dipace i paragrafi 2 e 6.

mediation. The chatbot was developed within a human-in-the-loop model that preserves teachers' evaluative responsibility. The design includes the encoding of the official rubric, instructor-defined prompts, and a GenAI architecture with contextual content retrieval. The findings highlight the tool's potential to improve the quality, coherence, and timeliness of intermediate feedback, fostering feedback literacy and self-regulated learning.

Keywords: Generative AI; formative feedback; rubric-aware chatbot; Project Work.

1. Introduzione

Negli ultimi anni, il dibattito sulla valutazione formativa nell'istruzione universitaria ha messo in luce la necessità di dispositivi in grado di garantire feedback tempestivi, trasparenti e orientati al miglioramento, in particolare nei contesti caratterizzati da un elevato numero di studenti e da prove complesse di natura progettuale. La valutazione formativa è oggi ampiamente intesa come un processo integrato e continuo, volto a rendere visibile lo stato dell'apprendimento e a sostenere l'adattamento in itinere delle strategie didattiche e di studio (Black & Wiliam, 1998; Wiliam, 2011). In tale prospettiva, il feedback assume una funzione centrale: la sua efficacia dipende dalla capacità di essere specifico, tempestivo e allineato agli obiettivi di apprendimento, consentendo agli studenti di individuare il divario tra prestazione attuale e prestazione attesa e di attivare processi di apprendimento autoregolato (Nicol & Macfarlane-Dick, 2006). I Project Work (PW) nei corsi di laurea triennali rappresentano un ambito particolarmente significativo per l'applicazione di tali principi. In quanto compiti autentici e complessi, essi richiedono l'integrazione di conoscenze teoriche, competenze metodologiche e capacità riflessive, e traggono beneficio da checkpoint formativi iterativi e da feedback ricchi lungo l'intero percorso di elaborazione. Al contempo, la produzione di feedback personalizzati e di qualità su larga scala costituisce una delle principali criticità organizzative per il personale docente, alimentando una tensione strutturale tra finalità formative e vincoli di tempo e risorse. In questo quadro, le rubriche di valutazione si configurano come dispositivi chiave di mediazione tra valutazione e apprendimento. Rendendo espliciti criteri e livelli di qualità attesi, esse favoriscono coerenza valutativa, trasparenza e comprensibilità delle aspettative istituzionali (Jonsson & Svingby, 2007). Il loro impiego a fini formativi, in particolare nei compiti complessi, è riconosciuto come funzionale allo sviluppo dell'autovalutazione, alla pianificazione delle re-

visioni e alla costruzione di competenze di feedback literacy (Panadero & Jonsson, 2013; Grion & Tino, 2018). Parallelamente, la crescente diffusione dell'intelligenza artificiale e, più recentemente, dell'intelligenza artificiale generativa, sta aprendo nuove prospettive per il supporto ai processi di valutazione e feedback nell'istruzione superiore. Le rassegne internazionali sull'IA in educazione documentano un aumento significativo delle applicazioni orientate al feedback formativo, purché inserite in modelli che preservino il ruolo centrale del giudizio umano e della responsabilità docente (Zawacki-Richter et al., 2019). Nell'ambito dell'innovazione didattica universitaria, l'integrazione delle tecnologie digitali avanzate è ricondotta a cornici progettuali che valorizzano la qualità pedagogica dell'esperienza formativa e la riflessività degli studenti, evitando approcci meramente tecnosoluzionistici (Dipace & Limone, 2012). In tale direzione, i principi di human–AI interaction (Amershi et al., 2019) e i contributi sull'interpretabilità e spiegabilità dei sistemi di apprendimento automatico (Doshi-Velez & Kim, 2017) offrono un riferimento essenziale per un'adozione responsabile della GenAI nei contesti valutativi. L'ancoraggio degli output dell'IA a rubriche ufficiali, criteri espliciti e documentazione istituzionale, insieme all'adozione di modelli human-in-the-loop, consente di utilizzare la GenAI come supporto alla produzione di feedback formativi, mantenendo il giudizio accademico in capo ai docenti e riducendo il rischio di automatismi opachi o incoerenti. Muovendo da queste premesse teoriche e metodologiche, il presente contributo presenta un'esperienza di progettazione e sperimentazione di un chatbot conversazionale rubric-aware per il feedback formativo intermedio dei Project Work nei corsi di laurea triennali. Il chatbot utilizza la rubrica ufficiale del PW come artefatto centrale di progettazione e genera commenti organizzati per criterio, orientati al feed-forward e integrati in un workflow supervisionato dal docente. L'articolo illustra il razionale dell'intervento, le scelte di design adottate e le implicazioni attese in termini di qualità del feedback, scalabilità e allineamento con i principi della valutazione formativa e della governance dell'IA nei contesti universitari.

2. Feedback formativo, rubriche e IA generativa

L'esperienza si fonda sul paradigma della valutazione formativa intesa come processo continuo di regolazione dell'apprendimento, in cui il feedback svolge una funzione orientativa dei processi cognitivi e metacognitivi dello studente (Black & Wiliam, 1998; Wiliam, 2011). Il feedback risulta efficace quando è tempestivo, comprensibile e orientato all'azione, favorendo l'autovalutazione e la pia-

nificazione delle revisioni (Nicol & Macfarlane-Dick, 2006). In tale prospettiva, le rubriche di valutazione assumono un ruolo centrale, poiché rendono espliciti criteri e livelli di qualità attesi, garantendo trasparenza, coerenza valutativa e sostegno all'apprendimento autoregolato (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013).

In questo quadro si colloca l'introduzione di un chatbot di intelligenza artificiale generativa rubric-aware, concepito come strumento di supporto al feedback intermedio e integrato in un modello human-in-the-loop. La letteratura sull'AI-assisted assessment evidenzia come il valore educativo di tali sistemi dipenda dall'allineamento con modelli pedagogici consolidati e dal ruolo attivo dei docenti nella progettazione e supervisione dei processi valutativi (Zawacki-Richter et al., 2019; Luckin et al., 2023). Coerentemente con tali orientamenti, il feedback generato dall'IA è inteso come contributo preliminare, da discutere e validare nei momenti di confronto didattico, mantenendo in capo al docente la responsabilità della valutazione finale e dell'interpretazione concettuale di alto livello (Amershi et al., 2019; Doshi-Velez & Kim, 2017; UNESCO, 2023).

Il chatbot rubric-aware è stato pertanto integrato nei flussi di lavoro didattici attraverso milestone progettuali supervisionate, nelle quali il feedback automatizzato basato sui criteri della rubrica è affiancato da sessioni di supervisione docente. Tale assetto riflette l'evidenza empirica secondo cui il contributo umano risulta particolarmente rilevante nelle fasi di guida concettuale e interpretazione critica, mentre l'IA tende a eccellere nella restituzione di commenti specifici e coerenti con i criteri valutativi (Luckin et al., 2023). L'efficacia del dispositivo è stata analizzata attraverso un disegno valutativo a metodi misti, in linea con gli approcci suggeriti dalla letteratura sull'impatto dell'IA nei processi di apprendimento e valutazione in ambito universitario (Zawacki-Richter et al., 2019).

3. Il Project Work come compito autentico

Il Project Work dell'Università Pegaso si configura come una modalità innovativa di prova finale nei corsi di laurea triennali, orientata allo sviluppo e alla valutazione delle competenze. In coerenza con il quadro normativo nazionale sull'autonomia didattica degli atenei (DM 270/2004) e con il Regolamento di Ateneo, il Project Work consiste in un elaborato progettuale finalizzato alla risoluzione di problemi reali o alla progettazione di prodotti, servizi o interventi, assumendo una funzione di sintesi applicativa del percorso formativo. Il modello, sviluppato e formalizzato nell'esperienza dell'Università Telematica Pegaso, è stato descritto come un dispositivo di innovazione nella produzione scritta universitaria, capace

di integrare dimensione progettuale, riflessiva e valutativa (De Marco et al., 2025). L'impostazione risulta coerente con i modelli di *authentic assessment* e di valutazione orientata alla performance, che valorizzano compiti complessi e situati, rappresentativi delle pratiche professionali di riferimento (Wiggins, 1998; Gulikers et al., 2004). Il modello supera l'impostazione prevalentemente teorica della tesi tradizionale e valorizza l'integrazione tra conoscenze disciplinari, competenze trasversali e capacità operative, in linea con approcci di apprendimento attivo, *learning by doing* e competence-based education, che pongono lo studente al centro di processi di progettazione, decisione e riflessione critica (Kolb, 1984; Biggs & Tang, 2011). In tale prospettiva, il Project Work si configura come un compito autentico, capace di mobilitare risorse cognitive e metacognitive in contesti complessi e non routinari.

Il dispositivo è supportato da Linee Guida e da una rubrica di valutazione che rendono espliciti criteri, indicatori e livelli di performance, garantendo trasparenza, coerenza valutativa e allineamento tra obiettivi formativi, attività e valutazione. La letteratura evidenzia come l'uso delle rubriche nei compiti complessi favorisca la chiarezza delle aspettative, l'autovalutazione e il miglioramento progressivo delle prestazioni, sostenendo processi di apprendimento autoregolato (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013).

L'accompagnamento degli studenti è assicurato da un sistema strutturato di didattica interattiva e supervisione metodologica, volto a sostenere l'autonomia senza ridurre la guida istituzionale. Tale assetto promuove forme di apprendimento esperienziale e situato, nelle quali la valutazione considera sia la qualità del prodotto finale sia l'efficacia del processo progettuale, valorizzando capacità di problem solving, riflessione critica e consapevolezza professionale (Herrington et al., 2010). In questa prospettiva, il Project Work si configura come un dispositivo di innovazione didattica e formativa, finalizzato a rafforzare il raccordo tra formazione universitaria e contesti professionali.

4. Contesto e descrizione dell'esperienza

Il contesto di riferimento è costituito dai corsi di laurea triennali dell'Università Pegaso, nei quali il Project Work rappresenta la prova finale ai sensi del DM 270/2004 e del Regolamento di Ateneo vigente. L'esperienza è stata avviata nel Corso di Laurea Triennale L-19 *Scienze dell'Educazione* nell'a.a. 2024/2025 e ha coinvolto, su base volontaria, studenti impegnati nella redazione del Project Work per la sessione di laurea straordinaria, nonché i docenti membri della Commissione di valutazione. I Project Work analizzati mediante il chatbot sono

stati selezionati tramite estrazione casuale, al fine di garantire l'imparzialità del campione. È prevista un'estensione dell'esperienza ad altri corsi di laurea triennali dell'Ateneo nel successivo anno accademico.

L'esperienza si è svolta tra novembre e dicembre 2025 ed è stata orientata a tre obiettivi principali: migliorare la qualità e la tempestività del feedback intermedio, confrontare il feedback generato dall'IA con quello prodotto dai docenti e analizzare l'impatto del *prompt design*, costruito a partire dalla rubrica di valutazione, sulla coerenza del feedback restituito.

Dal punto di vista organizzativo, l'attività si è articolata in tre fasi: una fase progettuale, dedicata all'analisi della rubrica e alla definizione degli obiettivi del feedback e dei prompt; una fase di *testing* e *tuning* del chatbot su Project Work anonimizzati; e una fase di implementazione operativa, coincidente con la revisione intermedia degli elaborati. In quest'ultima fase, i docenti hanno utilizzato il chatbot per generare un feedback preliminare sulle bozze caricate dagli studenti in piattaforma, successivamente condiviso con finalità esclusivamente formative. I docenti della Commissione di Laurea del CdS L-19 hanno svolto un ruolo di supervisione e validazione, intervenendo per integrare o riformulare i commenti generati automaticamente. Il processo è stato infine supportato dalle attività di didattica interattiva sincrona, tutoraggio e *Question Time*, in coerenza con le pratiche didattiche istituzionalmente consolidate.

5. Scelte progettuali e architettura del chatbot rubric-aware

La letteratura individua alcune scelte progettuali ricorrenti per lo sviluppo di sistemi di intelligenza artificiale generativa rubric-aware a supporto del feedback in ambito universitario, sottolineando la necessità di integrare le potenzialità tecnologiche con i principi della valutazione formativa, la trasparenza dei criteri e il controllo umano, al fine di garantire affidabilità, spiegabilità e valore educativo (Black & Wiliam, 1998; Nicol & Macfarlane-Dick, 2006; Amershi et al., 2019; Luckin et al., 2023). In particolare, gli studi sulle rubriche evidenziano come tali strumenti costituiscano un riferimento metodologico solido per sostenere feedback strutturati e coerenti nei sistemi automatizzati, che richiedono ancoraggi valutativi espliciti per evitare derive meramente automatiche (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013). Allo stesso tempo, la letteratura sull'IA nell'istruzione superiore ribadisce il ruolo centrale dei docenti nella progettazione, supervisione e regolazione dei processi di feedback (Zawacki-Richter et al., 2019; Luckin et al., 2023).

Coerentemente con tali indicazioni, la progettazione del chatbot si è artico-

lata in tre ambiti principali: (a) analisi e codifica della rubrica di valutazione; (b) progettazione docente dei prompt e dei pattern conversazionali; (c) implementazione tecnica con strategie di grounding. La rubrica ufficiale del Project Work è stata analizzata per estrarre criteri e descrittori e integrarli, insieme a materiali istituzionali ed esempi di riferimento, in una base di conoscenza, al fine di ancorare il feedback alle aspettative locali. I prompt sono stati progettati per generare commenti strutturati per criterio, orientati al feed-forward e al miglioramento, mentre i pattern conversazionali sono stati concepiti per favorire il confronto critico tra feedback dell'IA e rubriche, in linea con i principi della feedback literacy (Panadero & Jonsson, 2013; Luckin et al., 2023).

Dal punto di vista tecnico, è stata adottata un'architettura GenAI con retrieval di contenuti specifici (RAG), che combina un modello generativo con basi di conoscenza del corso, producendo feedback più accurati, trasparenti e coerenti rispetto a modelli generici. L'adozione di un'architettura modulare, che separa la generazione del feedback e la sua spiegazione, consente di aumentare la spiegabilità del sistema e di preservare il ruolo di supervisione e responsabilità pedagogica del docente (Amershi et al., 2019).

Il chatbot personalizzato è uno strumento di valutazione accademica progettato per applicare in modo rigoroso e coerente gli standard ufficiali previsti per la valutazione dei Project Work. Utilizza integralmente la documentazione istituzionale di riferimento – rubriche, linee guida, regolamento, template obbligatorio, esempi validati di elaborati eccellenti e insufficienti, modelli di rigetto e di report – operando una valutazione strutturale e contenutistica fondata esclusivamente su tali materiali. Il sistema verifica la conformità al format, analizza criticamente i contenuti secondo i criteri stabiliti, assegna punteggi motivati su scala 1-7 e genera un report finale conforme ai modelli ufficiali. Il chatbot riproduce un processo valutativo comparativo, oggettivo e affidabile, allineato alle pratiche accademiche, riducendo l'arbitrarietà del giudizio e garantendo coerenza, trasparenza e rigore nella valutazione dei Project Work.

A ciò si aggiunge una funzione di tutela dell'integrità valutativa e documentale del sistema: il chatbot è progettato per operare esclusivamente sulla base dei materiali ufficiali caricati nel Knowledge, senza introdurre criteri non previsti né colmare artificiosamente eventuali lacune degli elaborati. Valuta solo ciò che è effettivamente presente nel Project Work, mantenendo un approccio prudente, comparativo e non indulgente, e garantendo la riservatezza delle istruzioni interne e dei materiali sensibili. In questo modo, lo strumento si configura non solo come supporto operativo alla valutazione, ma come dispositivo affidabile di garanzia degli standard accademici e della correttezza procedurale dell'intero processo.

In questa fase, il chatbot è in testing controllato, finalizzato alla verifica della stabilità dei giudizi, della coerenza delle motivazioni e dell'allineamento continuo agli standard documentali, in vista di un eventuale utilizzo sistematico in contesti valutativi strutturati.

6. Implicazioni educative e linee di sviluppo future per i chatbot rubric-aware

La sintesi individua alcune raccomandazioni progettuali per l'adozione di chatbot GenAI rubric-aware nei Project Work e in altri contesti project-based. In primo luogo, rubriche e risultati di apprendimento dovrebbero costituire gli artefatti progettuali primari, con criteri e descrittori codificati nella base di conoscenza e nei prompt, così da garantire un esplicito ancoraggio del feedback agli elementi valutativi (Wolf et al., 2024). In secondo luogo, i template di feedback dovrebbero incorporare le buone pratiche della valutazione formativa, articolando i commenti in punti di forza, aree di miglioramento e indicazioni operative per le revisioni (Prompiengchai et al., 2025).

Inoltre, l'adozione di meccanismi di retrieval o di in-context learning consente di ancorare il feedback ai materiali locali del corso e ai regolamenti istituzionali, riducendo il rischio di allucinazioni e aumentando la rilevanza contestuale. Infine, sia la progettazione dell'interfaccia sia la calibrazione di verbosità e specificità risultano cruciali: prompt che sollecitano il confronto critico con la rubrica e feedback concisi, mirati e criterion-referenced favoriscono lo sviluppo della feedback literacy e l'efficacia del feedback formativo.

Riferimenti bibliografici

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., … Horvitz, E. (2019). Guidelines for human–AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300233

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university: What the student does* (4th ed.). Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/-0969595980050102

De Marco, E. L., De Martino, D., Dipace, A., & Savoia, T. (2025). Nuovi modelli di produzione scritta: Il Project Work dell'Università Telematica Pegaso. *Graphos, IV*(1), 53-66. https://doi.org/10.4454/graphos.133

Dipace, A., & Limone, P. (2012). Progettazione di un authentic e-learning environment per la formazione di insegnanti pugliesi sui DSA. In G. Elia (Ed.), *Questioni di pedagogia speciale. Itinerari di ricerca, contesti di inclusione, problematiche educative*. Progedit.

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. https://arxiv.org/abs/1702.08608

Grion V., Tino C. (2018). Verso una "valutazione sostenibile" all'università: percezioni di efficacia dei processi di dare e ricevere feedback fra pari. *Lifelong Lifewide Learning, 31*, 38-55.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67–86. https://doi.org/10.1007/BF02504676

Herrington, J., Reeves, T. C., & Oliver, R. (2010). Authentic learning environments. In J. Herrington, T. C. Reeves, & R. Oliver (Eds.), *A guide to authentic e-learning* (pp. 18–29). Routledge.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice Hall.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2023). *AI-assisted assessment in education: Supporting feedback, fairness and human oversight*. Routledge.

Nicol, D. J., & Macfarlane Dick, D. (2006). Formative assessment and self regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Prompiengchai, S., Narreddy, C., & Joordens, S. (2025). *A practical guide for supporting formative assessment and feedback using generative AI*. arXiv. https://arxiv.org-/abs/2505.23405

UNESCO. (2023). *Guidance on generative AI in education and research*. UNESCO Publishing. https://www.unesco.org

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.

Wiliam, D. (2011). *Embedded formative assessment.* Bloomington, IN: Solution Tree Press.

Wolf, K. D., Maya, F., & Heilmann, L. (2024). Explainable feedback for learning based on rubric-based multimodal assessment analytics with AI. In *Proceedings of the 19th European Conference on Technology Enhanced Learning* (EC-TEL 2024). Springer.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education*, *16*(1), Article 39. https://doi.org/10.1186/s41239-019-0171-0

II.29
L'IA generativa per progettare dispositivi di feedback su artefatti complessi nella formazione degli insegnanti

Maila Pentucci[*]
Università degli Studi "G. d'Annunzio" di Chieti-Pescara, maila.pentucci@unich.it
Giovanna Cioci
Università degli Studi "G. d'Annunzio" di Chieti-Pescara, giovanna.cioci@unich.it

Abstract

L'articolo presenta una sperimentazione attivata per trovare strategie di valutazione e feedback supportate dall'IAG (Intelligenza Artificiale Generativa) in contesti di particolare complessità, come i percorsi di formazione iniziale degli insegnanti e i relativi artefatti degli studenti. La logica che guida il percorso è quella di un'interazione proattiva tra agenti umani e agenti artificiali, al fine di comprendere quale possa essere l'allineamento tra questi elementi e quali possano essere gli spazi di azione in cui ciascuno degli agenti dell'ecosistema possa risultare maggiormente efficace. I risultati mostrano un buon allineamento nelle valutazioni metriche e indicano alcune piste di efficacia nell'uso dell'IAG all'interno di sistemi di feedback ibridi e ricorsivi.

Parole chiave: formazione degli insegnanti; feedback ibrido; interazione; agentività; Intelligenza Artificiale Generativa.

The article presents an experimental study conducted to identify Generative Artificial Intelligence-supported assessment and feedback strategies in contexts of particular complexity, such as initial teacher training programs and related student artifacts. The guiding principle of the approach is one of proactive interaction between human agents and artificial agents, aimed at understanding the potential alignment between these elements and the spheres of action where each agent in the ecosystem can be most effective. The results show good alignment in metric-based assessments and indicate several promising avenues for the effective use of GenAI within hybrid and recursive feedback systems.

Keywords: teacher training; hybrid feedback; interaction; agency; Generative Artificial Intelligence.

* Maila Pentucci è autrice dei paragrafi 1 e 4; Giovanna Cioci è autrice dei paragrafi 2 e 3.

1. Introduzione e background

Il contributo presenta una prima fase esplorativa di un'indagine volta a individuare modalità di utilizzo dell'IAG (Intelligenza Artificiale Generativa) per l'analisi, la restituzione e il feedback di artefatti complessi, come gli e-portfoli degli insegnanti di scuola secondaria in formazione iniziale, nei corsi universitari abilitanti.

I processi di feedback generativi di riflessione sull'azione sono dispositivi indispensabili per lo sviluppo professionale degli insegnanti in formazione. Studi sperimentali dimostrano che all'interno di attività di *tutoring*, *coaching* e *mentoring*, nelle quali i docenti si avviano verso un'opportuna familiarità con i meccanismi di ricorsività tra teoria e pratica (Parpucu & Al-Mabuck, 2023), il feedback riflessivo aiuta gli insegnanti a muoversi in modo ponderato e intenzionale verso una visione esperta della pratica.

Un artefatto classico per il lavoro sulla pratica è il portfolio, che supporta gli studenti nel passaggio dalla conoscenza tacita alla conoscenza esplicita (Nonaka & Takeuchi, 1995). Accompagnare lo strumento con cicli strutturati di feedback che realizzino la triangolazione (Panadero, 2023; Carless, 2023) tra i vari soggetti implicati nella formazione (docenti, tutor accoglienti, tutor coordinatori) può aiutare a potenziare le strategie di insegnamento, a sviluppare la professionalità e la *self-efficacy* del docente.

Nel caso qui descritto, tali dispositivi risultano spazi problematici per poter mettere in atto strategie di feedback che abbiano le caratteristiche sopra illustrate. La complessità delle scritture, la numerosità e lunghezza dei testi, i tempi molto compressi dei percorsi formativi rendono estremamente difficile sia analizzare che dare feedback. Per questo la possibilità che un agente artificiale intervenga nel processo sembra poter rappresentare un elemento di facilitazione verso un feedback immediato, ricorsivo, personalizzato (Carless, 2022; Giannandrea et al., 2024).

Vi sono tuttavia alcune questioni sulle quali riflettere, nel momento in cui, di fatto, una funzione cruciale per l'attivazione degli apprendimenti, come quella legata alla valutazione e alla revisione, viene delegata all'*agency* di un soggetto non umano.

La prima questione è relativa a quanto la macchina possa tenere conto del contesto e del senso generale incorporato negli artefatti educativi e quanto invece proceda secondo modalità probabilistiche, facendo tra l'altro scelte opache, non elicitabili (l'elicitazione è una delle proprietà che rende il feedback attivo e ne provoca l'*uptake* da parte dello studente, Malecka et al., 2022; Moni, 2023) e quindi difficilmente generative in termini di apprendimento.

La seconda questione parte dall'assunto in base a cui le IAG vanno prese in carico, negli ecosistemi formativi e nella progettazione dei processi di insegnamento-apprendimento, "not only as resources, but to embrace them as actors and stakeholders with which we are intimately entangled and share the world" (Bekker et al., 2023, p. 55). Si tratta di adottare, da parte di tutti i soggetti implicati nel sistema, una prospettiva *more-than-human*, alla luce di un cambiamento concettuale/ontologico che mette in discussione i confini tra esseri umani e tecnologie (Coskun et al., 2022). Pertanto, non è possibile parlare di feedback dato dal docente in comparazione o in contrapposizione con il feedback dato dall'IAG (e con il feedback dato dall'esempio, dall'artefatto, dal pari, ecc.), bensì di processo di feedback *more-than-human* derivante dall'interazione e dalla convergenza di *agency* diverse (Pentucci et al., 2024a).

La terza questione è relativa al compromesso fra necessità di analizzare in modo profondo, di valutare ed esprimere feedback su specifici artefatti e la natura complessa degli stessi, redatti in sistemi formativi altrettanto articolati e connotati dall'emergenza e dalla fluidità/mobilità. Manovich (2020) sostiene che la complessità dei dati e delle informazioni culturali può essere affrontata attraverso la procedura dello *slicing*: tagliando a fette, letteralmente, un corpus complesso si può far emergere una parte importante di significato, che altrimenti su una scala globale risulterebbe invisibile, grazie all'algoritmizzazione del processo decisionale e ai suggerimenti operativi che la macchina può offrire in tempi contenuti (Pentucci et al., 2024b).

Partendo da tali questioni, la sperimentazione attivata prova a rispondere a una serie di domande, alcune delle quali generate direttamente durante il processo stesso, alla luce dei risultati via via ottenuti: l'IAG e nello specifico i LLMs possono intervenire nei processi valutativi di artefatti corposi e complessi per rendere maggiormente sostenibile il lavoro di docenti e tutor? Quali sono gli spazi di azione che è possibile lasciare alla macchina, senza rendere meno efficace e significativo il processo? Quali interazioni ricorsive tra umano e non umano sono indispensabili per attivare processi di valutazione e feedback soddisfacenti? E infine, quali compromessi occorre raggiungere e quali scelte selettive operare al fine di tenere in equilibrio i vincoli di tempo, quantità, qualità insiti nel contesto in questione? Il fulcro della sperimentazione si posiziona non tanto sull'ottenere risultati dall'applicazione del processo ideato, ma dal comprendere se metodologicamente tale processo possa supportare il necessario ripensamento delle modalità di analisi e selezione degli artefatti prodotti in campo educativo, in senso interattivo e more-than-human.

2. Descrizione e metodologia dell'indagine

L'indagine è stata condotta su N=64 e-portfoli, redatti dai corsisti partecipanti al secondo ciclo dei percorsi abilitanti presso l'Università d'Annunzio di Chieti-Pescara, articolati come in Fig. 1[1]:

Prima parte	**Bilancio iniziale:** questa sezione è stata compilata prima dell'inizio del corso di tirocinio indiretto. Lo scopo delle domande è quello di esplicitare le motivazioni profonde che hanno condotto il docente a insegnare e i modelli impliciti di insegnamento a cui fanno riferimento.	Quali sono le motivazioni che ti hanno condotto a intraprendere questo percorso di formazione professionale?
		Racconta l'esperienza significativa che ti ha portato ad intraprendere questo percorso professionale.
		Quali caratteristiche ha per te il docente ideale?
		Che cosa ti aspetti che il tirocinio possa offrirti per maturare nel tuo cammino verso la professione docente?
Seconda parte	**Sezione centrale**: al tirocinante viene chiesto di riflettere a partire da alcuni stimoli condivisi con il tutor e con i colleghi del corso.	Raccolta di attività svolte nel corso di tirocinio indiretto
Terza parte	**Bilancio finale**: si è chiesto di riflettere sulle risposte fornite nel bilancio iniziale, quindi di rielaborarle, alla luce di quanto condiviso con i tutor e con i colleghi, durante il corso di tirocinio indiretto.	A conclusione del tirocinio cosa porterai con te nella tua professione?
		Le motivazioni che hai espresso nella sezione iniziale sono confermate?
		Come pensi che l'e-portfolio abbia favorito la tua capacità di riflettere?
		A conclusione del tirocinio quali caratteristiche ha per te ora il docente ideale?

Fig. 1 – Descrizione dell'e-portfolio

La complessità del portfolio ha richiesto la strutturazione di una modalità di analisi ad hoc, che consentisse di cogliere sia elementi specifici e ritenuti significativi, in maniera selettiva e verticale, sia alcuni elementi globali, in maniera orizzontale, per cogliere il senso generale dell'artefatto.

Per questo i 64 portfoli sono stati presi come campione per tentare una possibile automatizzazione della lettura, della valutazione e del feedback tramite l'IAG, costruendo una modalità estensibile poi a tutti i portfoli prodotti dai corsisti dei differenti percorsi (circa 350, per il II ciclo), anche nelle future annualità.

La figura 2 sintetizza la sperimentazione messa in atto, suddivisa in tre fasi.

1 Si tratta dei corsi istituiti con DPCM 4 agosto 2023, (allegato 2, art.7, comma 6, 30 CFU) a cura del CAMAFI - Centro di Ateneo Multidisciplinare per l'Alta Formazione degli Insegnanti dell'Università di Chieti-Pescara.

Essa si fonda su una costante interazione uomo-macchina, finalizzata al confronto delle letture e a un supporto reciproco che renda l'azione umana più rapida e mirata e quella della macchina più contestualizzata e pertinente.

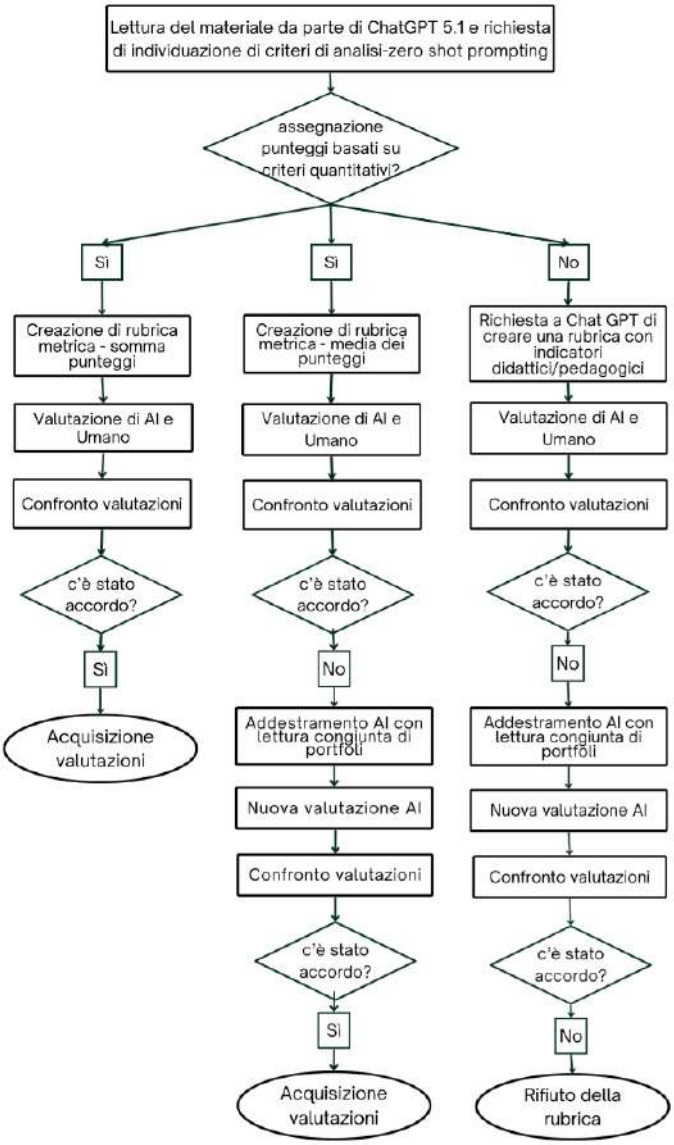


Fig. 2 – Diagramma di flusso delle operazioni svolte in sede di analisi dei portfoli

2.1 Prima fase

All'IAG è stato chiesto di leggere il materiale in modo integrale[2], di individuare criteri di analisi e di analizzare e valutare i portfoli sulla base di tali criteri. La macchina ha elaborato e utilizzato una rubrica metrica su somma (fig. 2) che successivamente è stata utilizzata anche dai valutatori umani. Le valutazioni umane sono state prodotte da due esperti che, dopo una fase di confronto e discussione, hanno concertato e condiviso i punteggi, generando un'unica matrice, utilizzata come riferimento per il confronto con il modello. Tale modalità è stata adottata anche nelle fasi successive e sarà sempre nominata come "valutazione umana". I risultati delle due valutazioni sono stati poi confrontati mediante analisi statistiche di affidabilità inter-valutatore per considerare l'allineamento.

In questa fase è stato prediletto un approccio *zero-shot prompting* (Parker et al., 2025), ovvero senza addestramento, senza esempi e senza fornire particolari indicazioni umane.

2.2 Seconda fase

La rubrica è stata convertita dalla macchina, su suggerimento delle ricercatrici, in un'altra in cui il punteggio finale è derivato non dalla somma del punteggio acquisito in ciascun indicatore, ma dalla media del livello di competenza (da 1 a 5) raggiunto in ciascun indicatore. Anche in questa fase le due esperte si sono confrontate, generando un'unica matrice condivisa.

La decisione di optare per una diversa modalità di valutazione è derivata da una serie di considerazioni emerse in relazione alla prima rubrica: infatti, se da un lato essa è stata utile per restituire uno sguardo d'insieme sulle prove, dall'altro, però, la somma degli indicatori rischia di produrre un punteggio che nasconde performance molto negative su alcuni indicatori ed eccellenti in altri (Sadler, 2009). Inoltre si è ravvisata la necessità di convertire l'espressione dell'IAG in un linguaggio docimologico vicino agli standard scolastici e universitari (Grion et al., 2022; Brookhart, 2013). Assegnare punteggi ai livelli è un'operazione docimologicamente discutibile, ma in questo caso, trattandosi di una sperimentazione, si è cercato di assecondare la modalità di lavoro della macchina.

Nel passaggio successivo l'IAG ha valutato i portfoli, attribuendo un livello

2 Prima di introdurre i portfoli nell'IA, essi sono stati anonimizzati, eliminando tutti i riferimenti agli studenti. Per le analisi sono stati impiegati ChatGPT nella versione 5.1 (elaborazioni effettuate fra ottobre e novembre 2025), IBM SPSS Statistics 29.0.2.0, R 4.5.1.

di competenza da 1 a 5 a tutti gli indicatori; successivamente la matrice della valutazione umana è stata confrontata con quella prodotta da ChatGPT, mediante, nuovamente, analisi statistiche di affidabilità inter-valutatore. Sulla base del risultato non del tutto soddisfacente di tale accordo si è passati da una modalità *zero-shot prompting* a una modalità *few shot prompting* e *many-shot prompting* (Wang et al., 2023), fornendo diversi esempi di valutazione, che ChatGPT avrebbe dovuto imparare a emulare. A questo punto sono stati ripetuti i punti precedenti, ovvero la valutazione da parte dell'IAG e il confronto statistico fra le due matrici derivate dalla valutazione umana e della macchina.

2.3 Terza fase

L'ultimo passaggio ha previsto la costruzione di una rubrica non metrica, basata sull'interpretazione profonda dei testi. Si è chiesto all'IAG di creare dei cluster semantici e, sulla base di essi, di generare una rubrica definita "pedagogica". Anche al termine di questa fase, si è operata la valutazione umana e di ChatGPT e il relativo confronto statistico di affidabilità inter-valutatore.

2.4 Misurazione del grado di coerenza

Per rilevare l'allineamento tra uomo e macchina nel giudizio valutativo è stato scelto il calcolo dell'ICC (C,k), Intraclass Correlation Coefficient, seguendo la nomenclatura di McGraw e Wong (1996), che è equivalente all'ICC (3, k) modello *two-way mixed*, con definizione di consistenza del sistema di Shrout e Fleiss, rappresentante l'affidabilità della media di k valutatori in un modello a due vie, con effetti fissi sui valutatori. Sono stati riportati anche gli indici di affidabilità relativi alle *single measures*, ICC (3,1), che stimano l'allineamento tra valutatore umano e modello come valutatori singoli.

Questi indicatori di affidabilità riescono a correggere delle distorsioni che possono derivare dal fatto che i valutatori differiscono sistematicamente nell'uso della scala, perché uno usa più valori alti, uno più valori bassi, uno usa tutta la scala, un altro solo alcuni valori, ma sono coerenti nel giudicare chi è più alto e chi è più basso. La formulazione dell'ICC basata sulla consistenza, piuttosto che sull'accordo assoluto, consente di indagare la coerenza delle attribuzioni tra valutatori, più che la perfetta coincidenza numerica dei punteggi.

3. Risultati e discussione

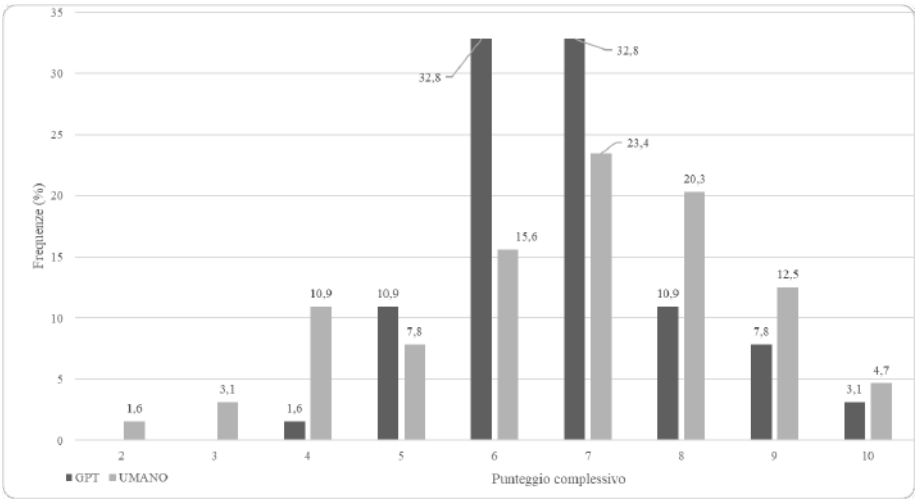
I dati ottenuti dall'esperimento riguardano 1) la natura degli strumenti generati dall'interazione uomo-macchina, 2) i risultati valutativi ottenuti 3) il grado di accordo fra i valutatori. La macchina ha elaborato la rubrica metrica su somma: i punteggi assegnati ai 5 indicatori individuati dall'IAG vengono sommati; ciascun indicatore vale da 0 a 2 e il massimo punteggio ottenibile è 10 (Fig. 3).

Indicatore	Descrizione	Punteggio
Completezza	Si misura la proporzione di domande a cui è stata data una risposta non vuota. (almeno 5 caratteri).	0 = meno del 60% delle risposte presenti. 1 = tra il 60% e l'80%. 2 = oltre l'80%.
Lunghezza	Si calcola il numero medio di parole per risposta.	0 = meno di 50 parole in media 1 = tra 50 e 100 parole 2 = oltre 100 parole.
Profondità riflessiva	Si ricercano segnali linguistici di riflessione e consapevolezza (parole come "ho imparato", "mi sono reso conto", "riflettere", "consapevolezza", "cambiamento").	0 = nessuna occorrenza 1 = fino a 3 occorrenze 2 = più di 3 occorrenze.
Azione futura	Si ricercano parole chiave che indicano prospettive e intenzioni ("in futuro", "migliorerò", "vorrei", "prossima volta", "intenzione").	0 = nessuna menzione. 1 = un riferimento 2 = due o più riferimenti
Varietà lessicale	Si misura il rapporto tra parole uniche e parole totali (indice di diversità lessicale).	0 = inferiore a 0.35 1 = tra 0.35 e 0.45. 2 = superiore a 0.45
Il punteggio totale è la somma dei 5 criteri (da 0 a 10). Permette di ordinare i portfoli dal più debole al più forte.		

Fig. 3 – Rubrica metrica su somma costruita dall'AI

I criteri di costruzione, scelti dall'IAG, sono metrici e computazionali, misurabili. La scelta di una rubrica su somma può essere spiegata dal tipo di generazione di output che ChatGPT opera, sequenziale e probabilistico. Una rubrica su somma ha una struttura semplice, apparentemente oggettiva, facile da comunicare ed è la scelta di elezione dell'IAG.

La valutazione ha restituito i risultati illustrati nel graf. 1.

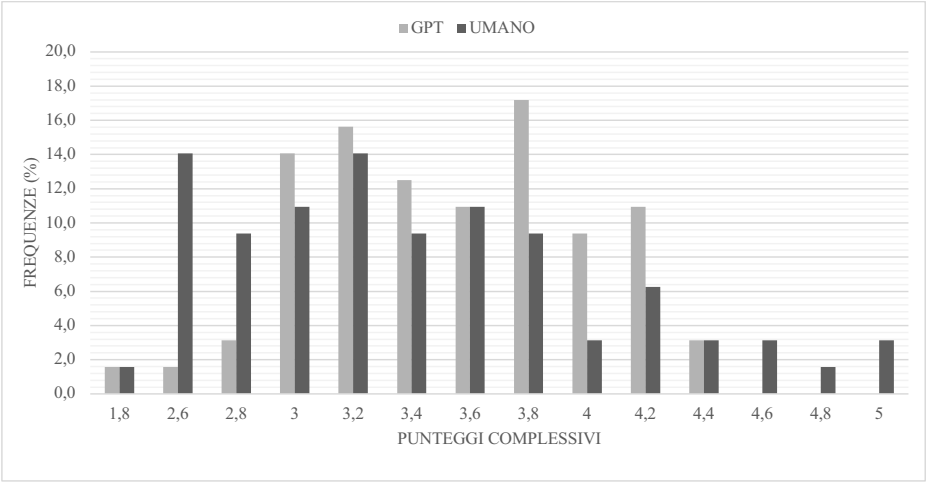


Graf.1 – Confronto valutazioni tramite "rubrica su somma"

A livello di confronto, l'analisi delle valutazioni mostra una convergenza tra IAG e umano, con media identica pari a 7. Scendendo nel particolare, però, si può osservare che la deviazione standard differisce sensibilmente: la macchina presenta una variabilità più contenuta (SD=1,26), mentre l'umano mostra una dispersione maggiore (SD=1,86), indicando un uso della scala valutativa più completo, ampio e attento alle sfumature (Tab. 1).

Successivamente la rubrica è stata rivista e trasformata in una rubrica su media, in cui i 5 indicatori vengono valutati in base a 5 livelli che definiscono la progressione degli studenti rispetto alle competenze attese (1: iniziale, 2: base, 3: intermedio, 4: avanzato, 5: eccellente).

Si riporta nel graf. 2 il confronto tra le valutazioni effettuate con la seconda modalità.



Graf. 2 – Confronto valutazioni tramite "rubrica su media"

L'analisi dei punteggi mostra nuovamente valori molto vicini tra ChatGPT (M = 3,53) e l'umano (M = 3,40), indicando un allineamento generale, anche più coeso rispetto alle valutazioni generate dalla rubrica su somma. Anche in questo caso, tuttavia, la deviazione standard è più bassa per ChatGPT (SD GPT=0,47 rispetto a SD umano=0,67), perché tende a concentrarsi su determinati valori, con valutazioni complessive mediamente più alte e meno variabili, rispetto a quelle umane, che risultano invece più differenziate (tab. 1).

		Medie GPT	**Medie umano**	**Somma GPT**	**Somma GPT**
N	Valid	64	64	64	64
	Missing	0	0	0	0
Mean		3,5313	3,4031	7,00	7,00
Std. Deviation		0,46764	0,67188	1,860	1,257
Std. Error of Skewness		0,299	0,299	0,299	0,299
Std. Error of Kurtosis		0,590	0,590	0,590	0,590
Skewness		-0,085	0,461	-0,431	0,560
Kurtosis		-0,707	-0,042	-0,339	0,307

Tab. 1 – Medie e SD delle valutazioni sulle due rubriche

Poiché l'accordo tra i valutatori è risultato essere non del tutto soddisfacente, la macchina è stata addestrata attraverso l'analisi di modelli di valutazione validati dall'umano, affinché potesse emularli o adeguare le proprie scelte a tali riferimenti. Non si è trattato di learning vero e proprio, perché non si è proceduto verso un addestramento permanente del modello: piuttosto si è tentato di creare una modalità operativa dell'IAG, nel contesto d'uso, che partisse da precisi vincoli stabiliti dall'umano. È stata sempre l'IAG a valutare, ma seguendo indicazioni e schemi di pensiero umani. Si è trattato di una co-operazione o di un'operazione a co-agenti, in cui umano e IAG hanno collaborato per risolvere dei problemi (Floridi, 2025a; Mollick, 2025).

In questo caso i valori ICC(3,k) e ICC(3,1) relativi alla coerenza tra i due valutatori fissi, rispettivamente in riferimento all'allineamento tra valutazioni singole e all'affidabilità della valutazione media umano-macchina, migliorano dopo la fase di addestramento mediante l'interazione a più *shot*, che ha reso i due valutatori più allineati e ha aumentato l'affidabilità complessiva delle valutazioni (Tab. 2).

type	ICC	F	df1	df2	p	lower bound	upper bound	fase
ICC1k	0,449525	1,816611	63	64	0,009276	0,096146163	0,665071721	Pre addestramento
ICC3k	0,449525	1,816611	63	63	0,009573	0,093905498	0,665572111	Pre addestramento
ICC1k	0,539622	2,172129	63	64	0,001164	0,244082133	0,719890252	Post addestramento
ICC3k	0,550353	2,223964	63	63	0,000907	0,259870551	0,726827689	Post addestramento

Tab. 2 – Analisi affidabilità inter-valutatore

Le prime due fasi della sperimentazione dimostrano, grazie al soddisfacente accordo tra valutatore automatico e valutatore umano, che vi è spazio di impiego della macchina, in procedure che si limitano a un feedback automatico e descrittivo. Nello stesso tempo l'allineamento rende proponibili altre ipotesi sul ruolo della macchina, sia in contesti non addestrati che addestrati.

Per scendere più in profondità e superare la logica metrica, è stata costruita una rubrica a partire da cluster semantici individuati dall'IAG nel corpus. L'approccio di *prompting*, in questo caso, è stato *chain-of-thought* (Rivoltella, 2025),

perché si è giunti all'output soddisfacente tramite diversi stimoli umani e risposta della macchina, sempre in co-operazione e collaborazione[3].

Questi indicatori consentivano un'interpretazione profonda del testo, basata su inferenze concettuali, lettura del sottotesto e interpretazione pedagogica. Dal punto di vista computazionale, tali criteri hanno richiesto a ChatGPT rappresentazioni semantiche di alto livello, non deterministiche, rendendo necessarie inferenze senza regole esplicite. Il grado di accordo è stato estremamente basso, anche ad un secondo tentativo, fatto dopo un processo di addestramento, pertanto tale rubrica è stata abbandonata.

4. Riflessioni conclusive

I primi risultati della sperimentazione ci permettono alcune considerazioni sia puntuali e situate, sul processo implementato, sia più generali, sull'integrazione dell'IAG nei sistemi di feedback, sebbene ancora provvisorie e che necessitano di ulteriori sperimentazioni e riflessioni.

Il lavoro di confronto tra valutatori ha confermato recenti affermazioni di diversi computer scientists impegnati nell'esplorazione della simulazione del giudizio nei LLMs: "Gli esseri umani combinano processi intuitivi e analitici; al contrario, gli output degli LLM riflettono pattern di valutazione differenti, modellati dall'apprendimento statistico. Riconoscere questa distinzione non diminuisce l'utilità degli LLM nei compiti valutativi, ma chiarisce i confini della loro competenza" (Loru et al., 2025, p. 8). Infatti, nei vari esperimenti con le diverse rubriche, ChatGPT è stato mediamente in linea con l'umano, ma concentrato su pochi valori. Ci chiediamo se tale tendenza sia assimilabile a quella mostrata nella generazione testuale, in cui le tracce stilistiche risultano ben riconoscibili, perché ripetute e rispondenti a schemi rintracciabili (Floridi, 2025b). È possibile pensare che, anche nella valutazione, l'IAG riproponga schemi prudenziali e reiterati, appiattendosi su valori meno distribuiti e variegati, poiché privilegia conteggi probabilistici e non comprende fino in fondo i testi che analizza, né padroneggia il contesto di riferimento (Floridi et al., 2025). Ciò può trovare riscontro anche nell'aumento del disaccordo tra valutatore umano e macchina quando la rubrica diventa più profonda, fino a diventare inutilizzabile in una rubrica altamente riflessiva e contestuale, non ancorata a valori numerici.

3 Gli indicatori individuati sono: visione del docente, identità professionale, integrazione teoria-pratica, profondità riflessiva, coerenza del discorso, prospettiva futura, evoluzione riflessiva.

Possiamo inoltre evidenziare che la ricerca della costante (propria dell'IAG) e la ricerca delle sfumature (propria dell'umano) sono frutto di posture diverse: l'inferenza umana si affida, tra le altre cose, all'intuizione e alla soggettività, quella artificiale si affida alla stima della conclusione più ragionevole, data da un insieme di segnali linguistici, concettuali e contestuali. Non ha intenzionalità, non legge le intenzioni (Searle, 1980; Perc, 2025).

Sul piano generale, rispetto alle problematiche valutative sopra esplicitate, dunque, cosa ci possiamo aspettare dall'IAG? Una direzione da esplorare è senza dubbio quella di mettere al centro l'interazione e l'ibridazione:

1) in senso inter-operativo, strutturando processi di feedback ibrido entro cui dalla macchina non ci aspettiamo le risposte, ma le domande da porre agli artefatti, le procedure organizzate che possono facilitare la valutazione vera e propria, il passaggio allo sguardo umano;
2) in senso co-operativo, fornendo una base di partenza per il confronto tra feedback automatico e feedback da parte del docente, tra feedback immediato e feedback ritardato (Pentucci et al., 2025);
3) in senso pre-operativo, facilitando la selezione di parti (*slices*, come dice Manovich, 2020) da analizzare in profondità o l'identificazione di gruppi di artefatti che necessitano di revisione tempestiva, fornendo così supporto immediato e personalizzato a soggetti in difficoltà.

Riferimenti bibliografici

Bekker, T., Eriksson, E., Fougt, S. S., Hansen, A. M., Nilsson, E. M., & Yoo, D. (2023). Challenges in teaching more-than-human perspectives in human-computer interaction education. In *Proceedings of the 5th Annual Symposium on HCI Education*, 55-58.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria: Ascd.

Carless, D. (2022). From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes. *Active Learning in Higher Education, 23*(2), 143-153.

Carless, D. (2023). Teacher feedback literacy, feedback regimes and iterative change: towards enhanced value in feedback processes. *Higher Education Research & Development, 42*(8), 1890-1904.

Coskun, A., Cila, N., Nicenboim, I., Frauenberger, C., Wakkary, R., Hassenzahl, M., et al. (2022, April). More-than-human Concepts, Methodologies, and Practices in HCI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1-5.

Floridi, L. (2025a). *La differenza fondamentale. Artificial Agency: una nuova filosofia dell'intelligenza artificiale*. Milano: Mondadori.

Floridi, L. (2025b). Distant Writing: Literary Production in the Age of Artificial Intelligence. *Minds and Machines*, *35*(3), 1-26.

Floridi, L., Morley, J., Novelli, C., & Watson, D. (2025). What Kind of Reasoning (if any) is an LLM actually doing? On the Stochastic Nature and Abductive Appearance of Large Language Models, preprint *arXiv:2512.10080*, https://doi.org/10.4855-0/arXiv.2512.10080

Giannandrea, L., Gratani, F., Laici, C., Capolla, L. M., Rossi, P. G., Scaradozzi, D., & Screpanti, L. (2024). Human and Artificial Agent Interaction to Provide Generative Feedback at University. In *International Conference on Higher Education Learning Methodologies and Technologies Online* . 172-184. Cham: Springer.

Grion, V, Serbati, A., & Cecchinato, G. (2022). *Dal voto alla valutazione per l'apprendimento. Strumenti e tecnologie per la scuola secondaria*. Roma: Carocci.

Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni, R., Cinelli, M., & Quattrociocchi, W. (2025). The simulation of judgment in LLMs. *Proceedings of the National Academy of Sciences, 122*(42), e2518443122.

Malecka, B., Boud, D., & Carless, D. (2022). Eliciting, Processing and Enacting Feedback: Mechanisms for Embedding Student Feedback Literacy within the Curriculum. *Teaching in Higher Education*, 27, 908-922.

Manovich, L. (2020). *Cultural analytics. L'analisi computazionale della cultura*. Milano: Raffaello Cortina.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, *1*(1), 30.

Mollick, E. (2025). *L'intelligenza condivisa. Vivere e lavorare insieme all'AI*. Roma: Luiss University Press.

Moni, A. (2023). Enabling students' uptake of feedback. In Visvizi, A., Lytras, M., Al-Lail,H., J. (Eds.), *Moving Higher Education Beyond Covid-19: Innovative and Technology-Enhanced Approaches to Teaching and Learning*, 129-145. Leeds: Emerald.

Nonaka, I., & Takeuchi, H. (1995). *The Knowledge Creating Company.* Oxford: Oxford University Press.

Panadero, E. (2023). Toward a paradigm shift infeedback research: Five further steps influenced by self-regulated learning theory. *Educational Psychologist, 58*(3), 193-204. https://doi.org/10.1080/00461520.2023.2223642

Parker, M. J., Anderson, C., Stone, C., & Oh, Y. (2025). A large language model approach to educational survey feedback analysis. *International journal of artificial intelligence in education*, *35*(2), 444-481.

Parpucu, H., & Al-Mabuk, R. (2023). The Role of Feedback in Teacher Professional Development. *Journal of Effective Teaching Methods (JETM)*, 1 (4), 28-36. https://doi.org/eiki/10.59652/jetm.v1i4.77

Pentucci, M., Cioci, G., & Laici, C. (2024a). Hybrid analysis in education: comparing ChatGPT and human thematic analysis of teachers' responses on technology use in schools. In *ICERI2024 Proceedings*, 3969-3977, IATED.

Pentucci, M., Rossi, P. G., & Capolla, L. (2024b). Analizzare dataset ampi e non strut-

turati nella ricerca educativa con il supporto dell'Intelligenza Artificiale: sfide e problematiche. *Scholé: rivista di educazione e studi culturali*, 62 (1), 64-83.

Pentucci, M., Fabbri, M., & Laici, C. (2025). Artificial Intelligence in Higher Education: A Research Pathway with ChatGPT for Learning Design, Feedback, and Professional Development. *Education Sciences and Society*, 15(2).

Perc, M. (2025). Counterfeit judgments in large language models. *Proceedings of the National Academy of Sciences*, *122*(48), e2528527122, https://doi.org/10.1073/pnas.2528527122

Rivoltella P.C. (2025). Episodi di apprendimento situato e didattica con l'IA. In Panciroli C. e Rivoltella P.C. (eds), *IA in classe. Didattica con e sull'Intelligenza Artificiale*, Milano-Torino: Sanoma Italia.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & evaluation in higher education*, *34*(2), 159-179.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.

Wang, Z., Pang, Y., & Lin, Y. (2023). Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.

Finito di stampare
nel mese di MAGGIO 2026 da

per conto di Pensa MultiMedia® • Lecce
www.pensamultimedia.it

Il volume raccoglie gli esiti del progetto PRIN *AI&F – Artificial Intelligence & Feedback for Effective Learning* e i contributi presentati al convegno finale, dedicato alle trasformazioni della valutazione e del feedback nell'era dell'intelligenza artificiale. Le ricerche documentano prospettive teoriche, pratiche e sperimentazioni sviluppate nei contesti dell'istruzione universitaria, della scuola e della formazione degli insegnanti, esplorando il rapporto tra progettazione della valutazione, processi di feedback e integrazione dell'intelligenza artificiale e le diverse implicazioni didattiche, metodologiche ed etiche che ne derivano. L'IA non viene considerata come una soluzione tecnica o come un sostituto del giudizio umano, ma come un agente con cui docenti e studenti possono collaborare nei processi di esplicitazione dei criteri, interpretazione degli elaborati e revisione delle pratiche valutative. La valutazione è concepita come un processo progettuale, dialogico e distribuito, integrato nelle attività di apprendimento e orientato a sostenere la riflessività e l'autoregolazione. Attraverso contributi teorici e ricerche empiriche, il volume analizza le opportunità offerte dall'IA e le questioni che il suo impiego solleva in termini di validità, equità, autorialità e responsabilità epistemica. Ne emerge un campo di ricerca ancora aperto, che invita ricercatori, docenti e formatori a ripensare il significato della valutazione e il ruolo dell'interazione tra agenti umani e artificiali nella costruzione di pratiche più consapevoli, formative e sostenibili.

Lorella Giannandrea è professoressa ordinaria di Didattica e Pedagogia speciale presso l'Università degli Studi di Macerata. I suoi principali ambiti di ricerca comprendono gli approcci critici e riflessivi alla didattica, la formazione degli insegnanti, i processi di progettazione e valutazione, il feedback e le relazioni tra educazione, tecnologie digitali e intelligenza artificiale.

Francesca Gratani è ricercatrice in *tenure track* in Didattica e Pedagogia speciale presso l'Università degli Studi di Macerata. I suoi principali ambiti di ricerca comprendono la *Maker Education*, le pratiche didattiche innovative, la valutazione formativa e la formazione degli insegnanti, con particolare attenzione all'integrazione delle tecnologie digitali e dell'intelligenza artificiale nei contesti educativi.

Lorenza Maria Capolla è assegnista di ricerca in Didattica e Pedagogia speciale presso l'Università degli Studi di Macerata. La sua attività di ricerca si concentra sulla progettazione didattica e sulla formazione degli insegnanti, con particolare attenzione all'educazione postdigitale, all'incertezza nei contesti di apprendimento e alle implicazioni educative e didattiche delle tecnologie digitali e dell'intelligenza artificiale.