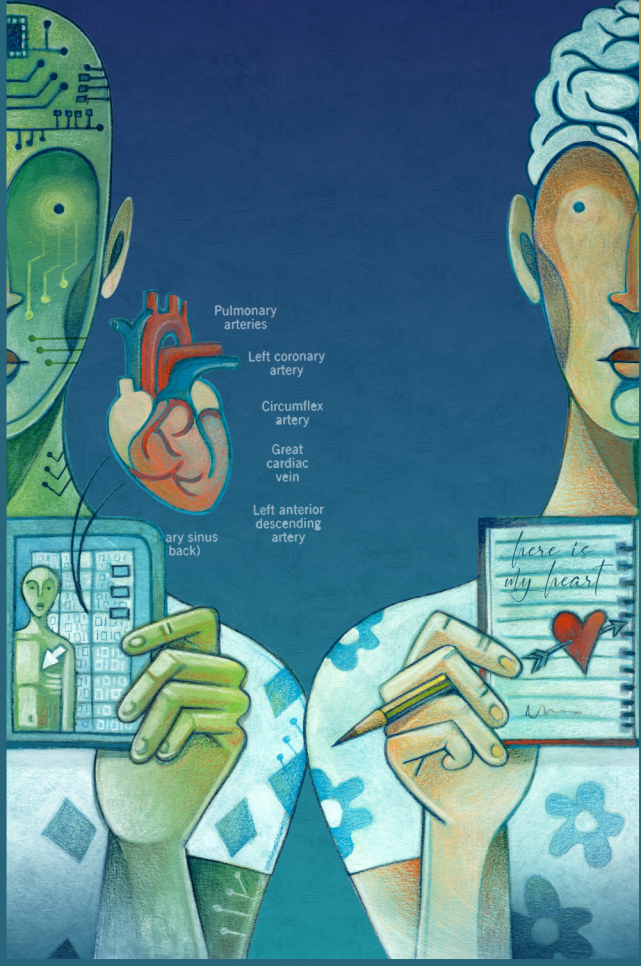
Lorella Giannandrea
Francesca Gratani
Lorenza Maria Capolla
[a cura di]



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

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.

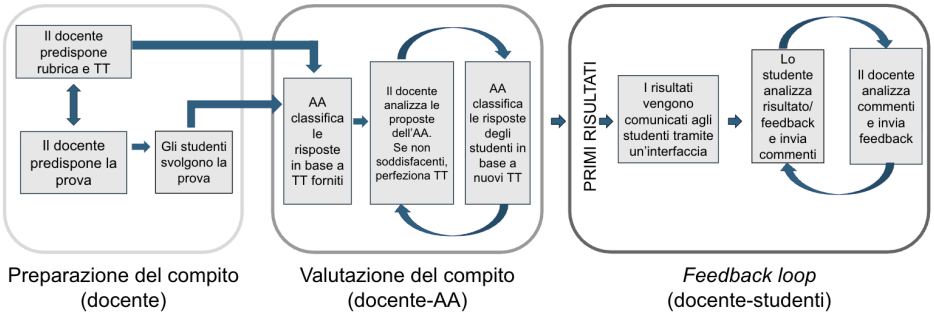


Fig. 1 – Il processo in sintesi (BERT)

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.

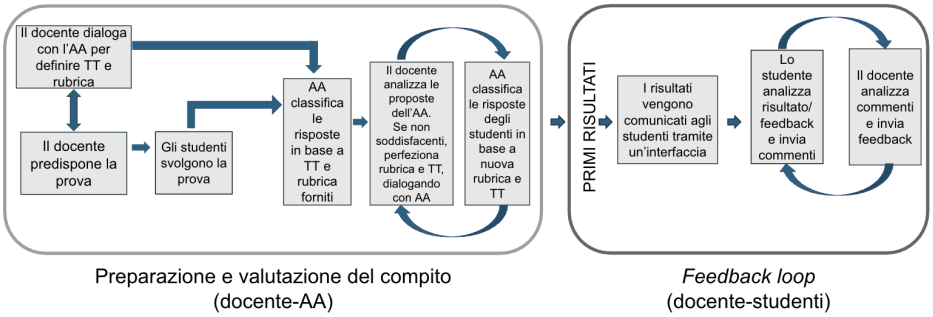


Fig. 2 – Il processo in sintesi (GenAI)

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.

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.