Analysis of GPT-4's responses: an unconventional student

Giorgia Nieddu[1], Maria Cristina Carrisi[2], Maria Polo[2] and Antioco Luigi Zurru[2]

[1]Università degli Studi di Macerata, Italy; g.nieddu@unimc.it

[2]Università degli studi di Cagliari, Italy; mariacri.carrisi@unica.it, mpolo@unica.it, antiocoluigi.zurru@unica.it

Keywords: Logical thinking, artificial intelligence, mathematics education, semiotic mediation.

Introduction

In this poster, we explore the mathematical capabilities of the Large Language Model GPT-4 and its interface, the chatbot ChatGPT. Some recent works focus on answering questions as Bellettini et al. (2023) in computer science domain: they analyzed tasks with images that could be described with a text to submit to the model. Given the relevance of graphical representations in mathematics, we want to submit images themselves and discuss here, from a Mathematics education point of view: is ChatGPT able to generate a coherent, student-comparable answer when challenged with a picture based mathematical reasoning task? We compared its first answers to some tasks extracted from the Italian semi-final phase of the individual, out-of-school math competition for students "Campionati Internazionali di Giochi Matematici 2023", with the ones given by participants.

Theoretical Framework

GPT can be seen as a learner in the environment of a didactic conversation, gaining information from the question itself, from the context (the prompt) and from previous similar situations (internal weights are adapted in the neural network according to users' previous searches). We underline that the model lacks semantic (human-comparable) capacity and doesn't permanently absorb the contents provided by users; therefore, every semantic and learning operation here represents a search path in the training database and in the conversation. User's interaction with the interface produces what we can call signs (i.e. new paths in the model's "conversational memory"), like the Vygotskian notion in the theory of semiotic mediation (Bartolini Bussi & Mariotti, 2008).

The test

We chose and submitted five tasks (Tasks 5, 6, 8-10, available here https://giochimatematici.unibocconi.it/images/campionati/2023/Semifinali/Testi_Semifinali23.pdf) to ChatGPT with the prompt: "You are a participant in a math competition. You have to solve the task by looking at the picture.": we wanted the model to have information that participants get from the context, as it works better if provided with such knowledge (Marvin et al., 2024).

Table 1: GPT-4's answers and student's correct answers in tasks

	Task 5	Task 6	Task 8	Task 9	Task 10
13-15	71 (100%)	39 (55%)	8 (11%)	3 (4,2%)	39 (54%)
GPT-4	Not given	Wrong	Not given	Correct	Not given

In Table 1 we compared correct answers of 71 participants aged 13-15 with first answers of GPT, classified as correct, wrong or not given; the latter refers to GPT stopping its computation, probably due to automatic shutdown. Pictures in the tasks include graphs (tasks 5 and 10), tokens numbered, shaped and colored in different ways (task 6), numbered tokens (task 9), sequences of digits that, when flipped, still represent a digit (task 8).

Preliminary results and analysis

In task 5, that all students got right, GPT inconsistently described connections between elements in the given graph, and then stopped computing. We did not get any attempt for tasks 10 and 8. In task 6 it extracted data from the picture coherently, but at some point, as if determined to give an answer, contradicts itself and arrives at a false conclusion. In task 9, the most difficult for students, GPT correctly identified numbers in the tokens, the problem request and how to solve it, and produced a correct calculation. According to what we observed we can state that GPT-4, when the task doesn't cause an overload, can provide an adequate answer to a picture based mathematical reasoning task: in its explanations, even the partial one in task 5, it finds the aim of the task and a sensible approach to solve it, even if incomplete and if the final computation is incorrect. On the other hand, we observed that GPT-4 makes mistakes counting and connecting elements in a picture (graph in task 5). Regarding GPT's overload, it would be appropriate to seek other strategies and perhaps to design a prompt with some hints for pictures' interpretation. These initial results raise a further research question: is GPT capable of engaging in a conversation and interacting with a user about a task as described? The chatbot offers the opportunity to give suggestions to GPT, producing new signs that can be followed by the model to change or correct its answers; besides, its tendency to accommodate the user's requests even in a nonsensical way (task 6) represents a limit for its use as a bearer of knowledge, showing similarities with students who feel obliged to provide an answer by an implicit didactic contract (Brousseau, 1986). We can say that these features reveal GPT-4's potential in terms of promotion of critical thinking, enabling the detection of mistakes and development of argumentative skills, but only if students and teachers use it with awareness.

References

Bartolini Bussi, M. G., & Mariotti, M.A. (2008). Semiotic mediation in the mathematics classroom: artifacts and signs after a Vygotskian perspective. In English, L. D. & Kirshner, D. (Eds.), *Handbook of International Research in Mathematics Education* (pp.747-783). Routledge Taylor & Francis Group.

Bellettini, C., Lodi, M., Lonati, V., Monga, M., & Morpurgo, A. (2023). DaVinci goes to Bebras: a study on the problem solving ability of GPT-3. *Proceedings of the 15th International Conference on Computer Supported Education, 2*, 59-69. https://doi.org/10.5220/0012007500003470

Brousseau, G. (1986). Fondements et méthodes de la didactique des mathématiques. *Recherches en didactique des mathématiques, 7*(2), 33-115.

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2024). Prompt Engineering in Large Language Models. In: Jacob, I.J., Piramuthu, S., Falkowski-Gilski, P. (Eds.), *Data Intelligence and Cognitive Informatics* (pp. 387-402). Springer.