DENNIS GABOR UNIVERSITY

THE USE OF ARTIFICIAL INTELLIGENCE IN SECONDARY AND HIGHER EDUCATION MATHEMATICS EDUCATION

Máté Bende*
Student, Alternative Economics High School, Budapest, Hungary
*Correspondence: bendemate17@gmail.com

Abstract

This study examines the educational potential of ChatGPT in mathematics learning by analyzing its problem-solving performance and students' perceptions of AI-generated explanations. The research evaluates ChatGPT's effectiveness on intermediate and advanced Hungarian high school mathematics exams and KöMaL competition tasks across subject areas such as algebra, geometry, and combinatorics. Results, manually scored using official answer keys, indicate high accuracy on exam-level problems and satisfactory performance on mid-level competition tasks, while difficulties emerge with the most advanced proofs and combinatorial challenges. A comparative analysis shows that task language (Hungarian vs. English) has minimal impact on performance. The second phase uses a questionnaire to explore students' attitudes toward ChatGPT, their preferences for AI-generated definitions, and their ability to identify machine-generated content. Findings suggest students often find AI explanations clear and cannot reliably distinguish them from human sources. Overall, the study highlights ChatGPT's promise as a supportive tool in mathematics education while emphasizing the need for critical use and pedagogical integration.

Keywords: artificial intelligence in education; mathematics problem solving; ChatGPT; student perceptions; AI-generated explanations

1. Introduction

The rapid development of artificial intelligence, especially large language models such as ChatGPT, is fundamentally transforming education. ChatGPT is capable of natural language communication, solving complex mathematical problems, and providing detailed explanations, thereby helping students gain a deeper understanding. In addition, AI offers personalized support to students, reducing the workload of teachers and making teaching more efficient.

The aim of my research is to examine ChatGPT's ability to solve mathematical problems based on intermediate and advanced level high school exams and KöMaL tasks. I analyze its performance in different subject areas (e.g., algebra, geometry, combinatorics) and whether the language of the task (Hungarian or English) influences its effectiveness. I evaluated the results manually, based on official answer keys.

The second part of the research is a questionnaire that maps students' habits and opinions regarding ChatGPT. I focus on examining the extent to which they prefer the definitions generated by the model and whether they recognize their machine origin.

My research is based on three main hypotheses. First, I assume that ChatGPT performs better on English-language math problems than on Hungarian-language ones, which may be due to differences in the model's language teaching ratios. Second, I believe that ChatGPT is highly reliable in solving math problems, regardless of their level of difficulty. Third, I hypothesize that students find the explanations generated by ChatGPT useful and often cannot distinguish them from content generated by human sources, which highlights the seamless integrability of AI content into education.

The novelty of the research lies in the fact that it provides a detailed performance analysis broken down by language and topic, while also integrating the student's perspective. Overall, ChatGPT has the potential to be a promising tool in mathematics education and can contribute to increasing the effectiveness of teaching.

2. Bibliography

Nothing demonstrates the importance of research into the relationship between large language models and mathematics better than the fact that the world's largest IT companies are investing billions of dollars in the development of various large language models, some of which specialize in mathematics. One of the largest closed-source models is Gemini Math-Specialized 1.5 Pro (Gemini Team Google (2024)), developed by Google, which was considered the most powerful large language model specialized in mathematics until August 2024, when the Qwen2-Math-72B-Instruct model was released (Qwen Team (2024)).

While Gemini is a closed-source model, the Qwen2 model, which surpassed it in August 2024, is open-source, meaning that anyone can freely access the entire model, every detail of which is public. As a result, Qwen2 is a particularly noteworthy model, as it not only performs excellently compared to the current state-of-the-art model, but also opens the door to a wealth of further development opportunities. But where does it lead when large language models far exceed the cognitive performance of the average person, not only in mathematics but also in everyday tasks? As LLMs become increasingly powerful and widely used, and we rely on them more and more, concerns arise about their impact on human cognitive abilities. According to an article by Richard Heersmink (Heersmink (2024)), LLMs may affect our cognitive abilities by taking over tasks that require critical thinking, such as solving mathematical problems, proving theorems, or even writing a simple email. Heersmink argues that if we rely too heavily on LLMs, we may lose certain skills, similar to how using GPS can reduce our spatial orientation.

We must not forget the role of artificial intelligence in education. In his article (Lo (2023)), C. K. Lo analyzed 50 scientific papers on the relationship between AI and education. In his work, he reports on the different ways in which AI is used for education in different countries around the world, the problems that arise, and the solutions that can be applied. One of the most fundamental critical questions is whether AI is capable of generating reliable, correct answers to questions at all. I address this question in my thesis, in which I show that ChatGPT, currently the best-known publicly available artificial intelligence, has an excellent understanding of middle and high school mathematics, and even, to a certain extent, more serious competitive tasks. However, in cases where AI does not perform well in a given area, it is possible to fine-tune it for a specific field of science (Google Cloud: Dialogflow (2023); Topsakal and Topsakal (2022)). If, on the other hand, AI excels in a particular field of science, such as mathematics, it can be used in many ways in education, for example, it can help students learn by explaining the curriculum, checking their work, pointing out and explaining mistakes, and providing correct answers, all while being available at any time. It could even make teachers' work easier, for example by coming up with assignments or generating comprehensive teaching materials and notes for a given topic (Lo (2023)).

3. Does language matter?

It is also interesting to examine whether the language of the task matters in the case of a large language model, since in theory it should be able to translate the received text independently. Based on my previous interactions, I found that the model is somewhat less accurate in Hungarian and sometimes phrases things a little awkwardly, but I assumed that this difference would be less significant in the field of mathematics, as the world of numbers is language-independent. To test this, I tested the model's performance on the 2024 advanced level mathematics exam questions, first in Hungarian, then I had it solve the same questions translated into English in a new window. The result: the performance proved to be almost identical. It scored only 1 point less on the Hungarian problems than on the English ones (see Table 1).

Table 1. Partial results achieved by GPT-4o on English and Hungarian test sheets.

Task number	Maximum score	4o Hungarian	4o English	Absolute difference
1a	6	6	6	0
1b	5	3	4	1
1c	3	3	3	0
2a	4	4	1	3
2b	4	4	4	0
2c	3	3	3	0
3a	5	2	2	0
3b	3	3	1	2
3c	4	4	4	0
4a	4	1	4	3
4b	6	4	3	1
4c	4	4	4	0
5a	3	1	1	0
5b	3	3	3	0
5c	5	3	5	2
5d	5	0	2	2
6a	5	1	2	1
6b	9	8	8	0
6c	2	1	2	1
7a	5	5	5	0
7b	3	3	3	0
7c	8	2	3	1
8a	3	3	2	1
8b	8	0	0	0
8c	5	2	1	1
9a	5	5	5	0
9b	5	5	3	2
9c	6	3	3	0
Total	131	86	87	

4. Mathematics Final Exam

The first test series I used to evaluate ChatGPT's performance was the mathematics final exam, which all students wishing to continue their studies must take at the end of their high school education.

Students can take the mathematics exam at two different levels: intermediate and advanced. The intermediate exam basically consists of two parts: in the first half of the exam, they must solve simple short-answer questions (questions 1-12), for which they can earn 2-3 points, and in the second half, they must solve explanatory calculation questions (questions 13-18), which are more complex, consist of several parts, and require more detailed explanations during grading. Students only have to solve two of the last three tasks in the second block. The advanced level consists of one part of the final exam. Students must complete nine complex essay questions, of which they only need to answer four of the last five.

In this chapter, I evaluated five years' worth of final exam questions, both at the intermediate and advanced levels, for a total of more than 270 questions. I solved the tasks using both the GPT-4o and GPT-o1 models and examined the answers. I read through the solutions and thought processes generated by the models one by one and then scored them manually based on the official answer key. I collected these in an Excel table (see Table 2).

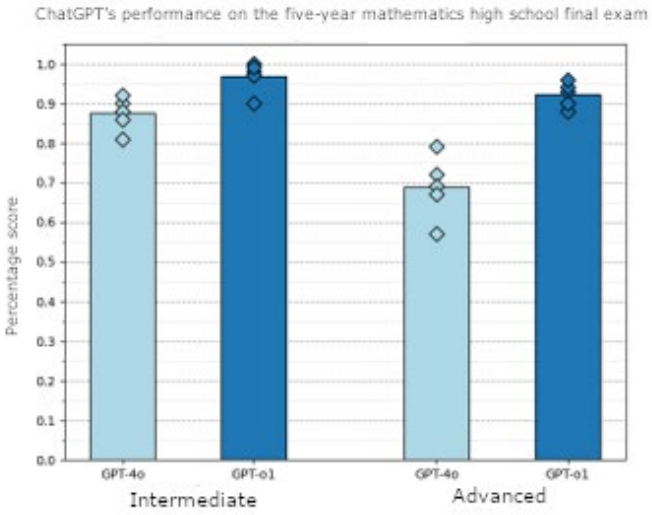
Table 2. In Excel, in addition to the achieved and maximum scores, I also collected data on which topic the task belonged to, what the final results were (and whether they matched), and the type of errors.

%%	ex num	answer ori	answer key	type	max	achived	final score	error1	error2
	2b	2049 integer solutions	2020 integer solutions	algebra	6	3	no	ignores	
	3a	260,000 people seeking jobs	260,000 people	algebra	3	3	yes	-	
	3b	67.85% of the working-age population	0,68	algebra	4	4	yes	-	
	3c	Numbers: {1, 2, 3, 4, 5, 100, 200}, Mean = 45, Median = 4	Numbers: {1, 2, 3, 4, 5, 6, 21}, Mean = 6, Median = 4	statistics	3	3	yes	-	
	3d	Mean can be skewed by outliers; conclusion is incorrect	Correct	statistics	2	2	yes	-	
	4a	Center: (2.8, 5.6), Area: 78.32 cm^2	Center: (2.8, 5.6), Area: 78.4 cm^2	geometry	7	7	yes	-	
	4b	21 cm^2	9 cm^2	calculus	7	3	no	calculated with an incorrect number	
	5a	126 ways	15 ways	combinatorics	5	2	no	wrong solution path	
	5b	Probability = 14/75	Probability = 1/60	probability	5	2	no	misinterpretation	

Summarizing the results, we can observe the performance shown in Figure 1. It can be said that at the intermediate level, both the GPT-4o and GPT-o1 models performed well, with average performances of 88% and 96%, respectively. At the advanced level, however, the performance of GPT-4o dropped to 69%, which is still an excellent level, while the GPT-o1 model achieved a performance of over 90% even at the advanced level. Basically, it can be observed that both models perform very well on intermediate-level simple short-answer tasks, solving them excellently with only a few misunderstandings. In the case of more

complex, explanatory tasks, there are already significant differences between the performance of the two models at the intermediate level, but at the advanced level, this difference is already clearly visible.

Figure 1. ChatGPT's performance based on the 5 years of data I evaluated. Overall, both models performed quite well at the intermediate level, but at the advanced level, GPT-4o performed much weaker than the GPT-o1 model.



5. KöMaL

This paper examines the solubility of KöMaL's monthly math problems using the ChatGPT-o1 model. KöMaL problems are divided into four difficulty levels: categories K, C, B, and A. K problems are for beginners, C problems are for intermediate learners, while B and A problems are serious competition problems, the latter even preparing students for international competitions. Although C tasks are related to the school curriculum, they can be more difficult than the advanced level school-leaving exam, as participants have up to a month to solve them, even in teams.

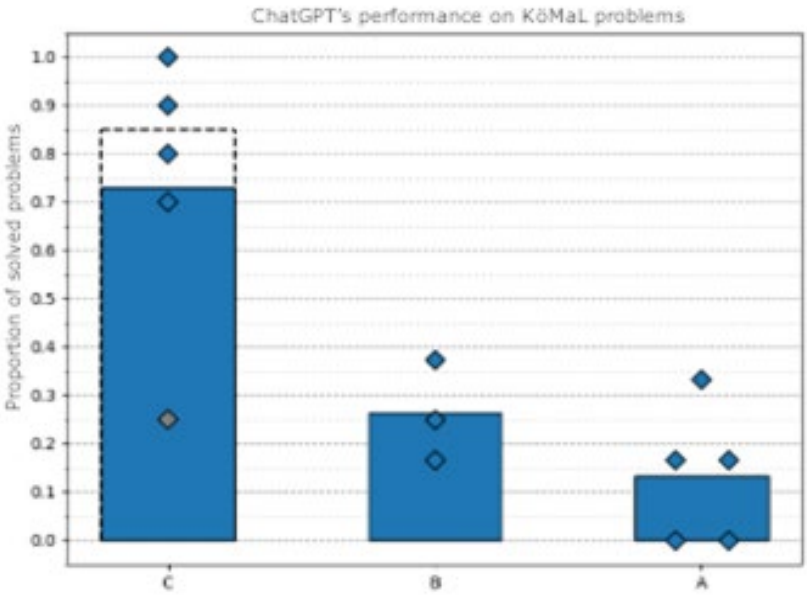
The GPT-4o model previously performed poorly on the advanced level exam, while the GPT-o1 model performed well, so the analysis was only performed on the latter model. The study only covers tasks marked C, B, and A, which are intended for upper grades.

During the evaluation, it was not only complete accuracy that counted: if a solution contained the main line of reasoning or only minor errors (e.g., calculation errors), I accepted it. If the model reached an important step but did not completely solve the task, I gave half a point—but this was rare. Typically, the model either solved the task nicely or gave a completely wrong answer. (See Figure 2.)

During the analysis, I examined 5 months of C-level tasks (25 tasks), 3 months of B-level tasks (24 tasks), and 5 months of A-level tasks (15 tasks). The ChatGPT-o1 model performed relatively well on the C-level tasks, solving an average of 4.5 tasks, although a weaker set of tasks lowered the average. A significant decline was observed in the B-level tasks: the model was only able to solve 25% of the tasks . In the case of tasks marked A, it solved only one task, while in the other cases it made attempts that were at best rudimentary.

Based on this, it can be concluded that the model is not suitable for tasks marked with A, and it also struggles with those marked with B – the limits of ChatGPT probably lie between tasks marked with C and B. A typical error was that it tried to answer geometric problems too rigidly using coordinate geometry methods, which in many cases made the solution more complicated. In the case of proofs, it often examined the correctness of a statement only through a specific example, rather than in a general way.

Figure 2. Performance of the o1 model on KöMaL tasks

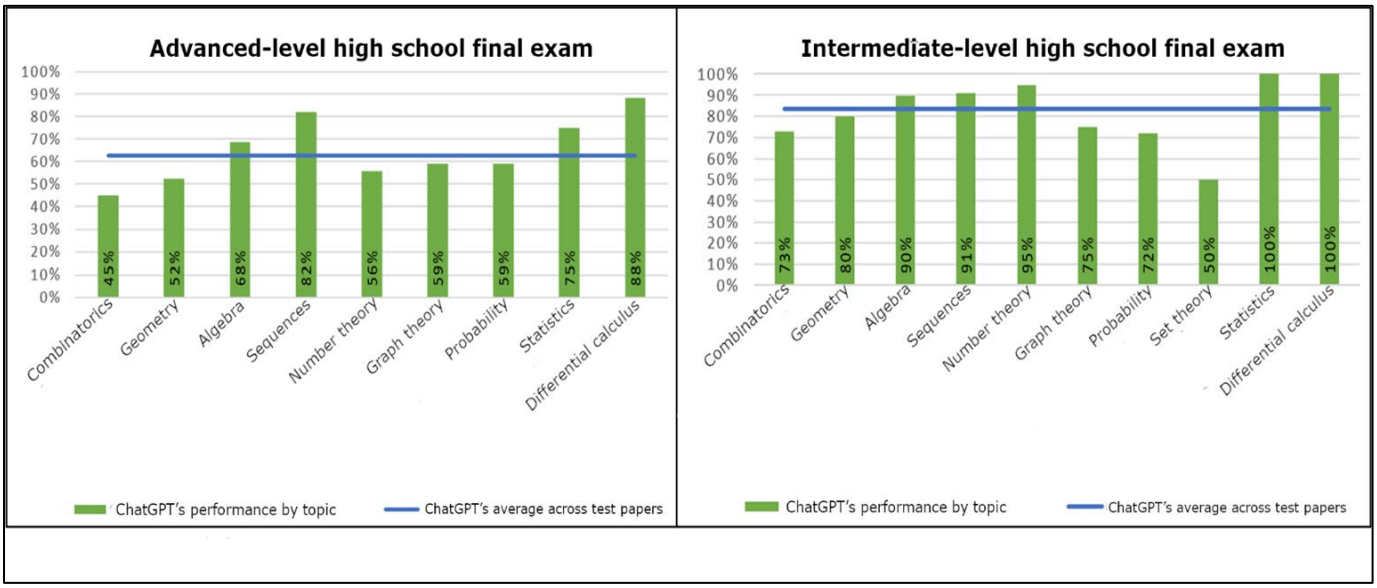


Although I do not illustrate this in my thesis, it is important to mention that several AI studies use a method whereby the model is run several times on the same task, and the results are then evaluated retrospectively to see if there is a usable solution among them. It is conceivable that a useful idea may emerge when a problem is generated 10 times. However, this type of use requires more advanced mathematical knowledge, as the user must be able to filter out bad solutions and recognize valuable elements. This is time-consuming and energy-intensive and requires critical thinking, but an experienced competitor can still benefit from using AI.

6. Results by topic

I grouped the tasks by topic, which makes it possible to examine how the model performs in each topic (see Figure 3). Overall, it can be observed that the successful completion of the tasks is influenced by the subject area from which the questions originate, as there are tasks at both intermediate and advanced levels on which the model performs above or below average. , for example, excels at solving tasks related to sequences and differential calculus, while combinatorics is clearly one of its weak points.

Figure 3: Performance of the 4o model at intermediate and advanced levels.



7. Questionnaire

The questionnaire used in the research was designed to provide a comprehensive picture of students' experiences and opinions regarding ChatGPT. The questions were divided into three main sections. In the first section, I asked respondents about their grade level, attitude toward mathematics, and ChatGPT usage habits. The second section examined the comprehensibility and recognizability of the definitions generated by the model, focusing in particular on the extent to which students preferred the explanations provided by ChatGPT ( ) and whether they were able to distinguish them from other sources. The third part asked about the future role of artificial intelligence, primarily its impact on learning and mathematics education.

Table 3. Distribution of students based on ChatGPT usage. 100% corresponds to 107 completions.

Educational level	Yes	No
University: Bachelor's level	42,11%	57,89%
University: Master's level	57,14%	42,86%
Upper secondary school: Grades 11-13	17,65%	82,35%
Lower secondary school: Grades 7-10	28,13%	71,88%

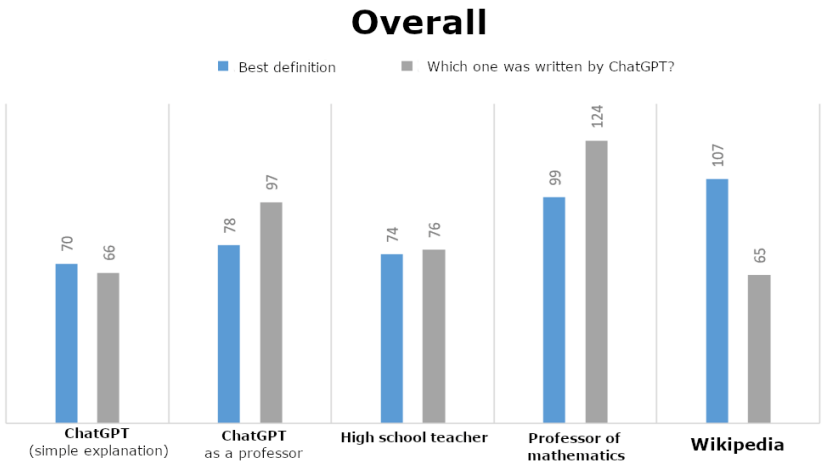
The questionnaire included various response formats, such as single and multiple choice options, as well as rating scales, to allow respondents to express nuanced opinions. A total of 107 people completed the questionnaire, including a mix of high school and university students.

8. Examination of definitions

In this section, I asked Turing test-like questions: I gave five different definitions for each of four concepts (geometric series, logarithm, limit, and derivative). The five definitions came from the following sources: a high school teacher, a university mathematics professor, Wikipedia, and two ChatGPT-generated answers, one of which was a simple definition and the other a professor-style definition. I asked respondents to first indicate the definition that was most understandable to them, and then the one they thought was created by ChatGPT.

Based on the aggregate results (see Figure 4), students found the definitions written by Wikipedia and the mathematics professor to be the best. Interestingly, however, most recognized the definitions written by the mathematics professor as originating from ChatGPT. Furthermore, I used matrices to examine who the respondents attributed the definitions they found most understandable to. The results were completely representative: for example, those who chose the professor's definition as the most understandable rated the other options proportionally in line with the other respondents.

Figure 4. Aggregate results: Best definition and ChatGPT recognition



The figures clearly show that upper-level students found the definitions generated by ChatGPT to be just as understandable as those from any other source; in fact, when it came to the concept of a limit, most voted for the professor-style ChatGPT definition. The results were different among university students: for them, the definitions provided by the mathematics professor and Wikipedia were the most understandable, and they were less likely to choose ChatGPT's explanations as the best. When students had to choose which definition was created by ChatGPT, my hypothesis was confirmed: they were unable to consistently distinguish ChatGPT's definitions from the others. In the case of geometric sequences, both high school and university students identified the professor-style ChatGPT explanation, but for other concepts, their responses varied greatly, and they mostly thought that the definitions written by the mathematics professor were ChatGPT-generated.

9. Testing hypotheses

I also examined the hypotheses formulated during the research in detail. One preliminary assumption was that language plays a significant role in the performance of the model. The test results did not confirm this: the performance of the ChatGPT-o1 model was essentially language-independent, showing similar effectiveness in both Hungarian and English tasks. The second hypothesis was that ChatGPT can be successfully applied to solve intermediate and advanced level high school exam tasks, which was clearly confirmed, as the model performed with over 90% accuracy at both levels. The third hypothesis examined was the extent to which students can recognize and evaluate the explanations provided by ChatGPT. The results showed that high school students in particular did not show a clear preference and were often unable to distinguish between explanations generated by artificial intelligence and those provided by teachers or professors. This suggests that the answers provided by ChatGPT have already reached a level of content and language that can be incorporated into formal education without deviating distractingly from traditional sources.

10. Concluding thoughts

In my analysis, I examined the performance of OpenAI's models, which are the best known and most powerful models: GPT-4o and GPT-o1. Based on my personal experience and an evaluation, I have confirmed that these new models perform at almost the same level in mathematics, regardless of language. I evaluated the performance of the models on intermediate and advanced level mathematics exam questions from the past five years, and also examined them on KöMaL tasks.

The development of different generations of ChatGPT shows spectacular progress in the field of mathematical problem solving. The latest GPT-4o model would now achieve excellent results on both intermediate and advanced high school exam problems, while the earlier GPT-o1 model—which, according to Terence Tao, is at the level of an average master's student—would perform above 90% even at the advanced level. In my research, I analyzed the performance of the GPT-o1 model not only on high school exam problems, but also on KöMaL competition problems. According to my results, the model was able to solve the C-level, intermediate-level problems satisfactorily, but performed less well on the B-level and especially the A-level, advanced-level examples. However, this is still a significant improvement, considering that previous ChatGPT models often performed even basic calculations, such as multiplication, incorrectly. The trend in the development of artificial intelligence is clear: models are becoming increasingly accurate and rich in knowledge. OpenAI's current development, the GPT-o3 model, is expected to have PhD-level knowledge, which could further strengthen the role of such models in the teaching and application of mathematics.

With this research, we have taken the first critical step towards making AI systems truly integrable into education. This requires a thorough understanding of the possibilities and limitations of artificial intelligence, without which we would not be able to use them properly. The results show that these models, especially the o1 model, perform excellently not only in solving simple tasks but also in more complex ones. Therefore, if a student needs help, these AI models can provide competitive support, which can make them a useful tool in education. The questionnaire responses also showed that the model performs well in defining abstract concepts. Many found its explanations to be the most understandable, and these definitions did not seem to be "machine-generated," which is particularly positive feedback. In addition, it was found that high school students are generally open to

ChatGPT being used in math classes. This may offer teachers a new opportunity to use artificial intelligence as a support tool in their classes, helping students understand and improve in the subject.

It would be worthwhile to continue this research in several directions in the future. Performance could be compared with other artificial intelligence models, such as Claude.ai, Gemini 2.5 Pro, Qwen2, or ChatGPT o4-mini-high, to find out which model performs better with different types of math problems. With greater computing capacity, it would even be possible to build a target model similar to AlphaGeometry, a system specialized for KöMaL competitions, which would likely achieve better results on both B and A tasks.

The questionnaire could also be further developed: in a new section, we could show students detailed, step-by-step solutions to specific mathematical problems and then ask them to evaluate their comprehensibility and usefulness. This would provide more accurate feedback on how effective the explanations of the different models are in practice. Future studies could even be conducted in a more formal setting, in the form of structured, multi-hour sessions, using pre- and post-tests.

References

1. Dinh, T. A., Mullov, C., Bärmann, L., & et al. (2024). Sciex: Benchmarking large language models on scientific exams with human expert grading and automatic grading. (https://arxiv.org/abs/2406.10421).
2. Gemini Team Google. (2024). Gemini: A family of highly capable multimodal models. (https://arxiv.org/abs/2312.11805)
3. Google cloud: Dialogflow. (2023).(https://cloud.google.com/dialogflow).
4. Heersmink, R. (2024). Use of large language models might affect our cognitive skills. NatureHuman Behaviour, 8, 805–806.
5. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation,9(8).
6. Imani, S., Du, L., & Shrivastava, H. (2023). Mathprompter: Mathematical reasoning using large language models. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 5, 37–42.
7. Lo, C. K. (2023). What is the impact of chatgpt on education? A rapid review of the literature. Education Sciences, 13(4).
8. OpenAI. (2024). Gpt-4 technical report. (https://arxiv.org/abs/2303.08774).
9. Qwen Team. (2024). Qwen2 technical report. (https://arxiv.org/abs/2407.10671).
10. Tao, T. (2024a, June 8). AI will become mathematicians' co-pilot. Retrieved from https://www.scientificamerican.com/article/ai-will-become-mathematicians-co-pilot/ .
11. Tao, T. (2024b, October 4). Interview in the Atlantic. Retrieved from https://www.theatlantic.com/technology/archive/2024/10/terence-tao-ai-interview/680153/.
12. Topsakal, O., & Topsakal, E. (2022). Framework for a foreign language teaching software for children utilizing AR, voicebots, and ChatGPT (large language models). The Journal of Cognitive Systems, 7(2).
13. Trinh, T. H., Wu, Y., Le, Q. V., He, H., & Luong, T. (2024). Solving olympiad geometry without human demonstrations. Nature, 625, 476–482.
14. Vaswani, A., Shazeer, N., Parmar, N., & et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 6000 - 6010.