

Research Article

Examining the potential and pitfalls of AI in problem solving

Yasin Memiş

Niğde Provincial Directorate of National Education, Türkiye (ORCID: [0000-0001-6961-3452](https://orcid.org/0000-0001-6961-3452))

The integration of artificial intelligence (AI) into mathematical problem-solving has shown significant potential to enhance student learning and performance. However, while AI tools offer numerous benefits, they are prone to occasional conceptual and arithmetic errors that can mislead users and obscure understanding. This research examines such errors to improve the role of AI in solving mathematical problems. The study particularly assesses the abilities of AI tools—ChatGPT-4, Gemini, and CoPilot—in addressing proportional reasoning errors commonly made by students. ChatGPT-4 achieved the highest accuracy rate among the tested tools, correctly answering 10 out of 14 questions. Additionally, ChatGPT-4 provided more detailed explanations in its responses, with a higher word count compared to the other tools. However, all tools replicated certain errors commonly made by students on specific reasoning questions. In conclusion, while AI tools hold promise for enhancing mathematics education, they still have limitations. Improving AI's contextual understanding and problem-solving adaptability could lead to the development of more robust educational tools.

Keywords: Artificial intelligence; Mathematics education; Problem-solving; Proportional reasoning; Large language model

Article History: Submitted 8 November 2024; Revised 18 February 2025; Published online 11 March 2025

1. Introduction

AI is garnering significant interest across various fields, including education. Over the past 20 years, 64% of studies on the use of AI in education in Türkiye have been conducted in the last five years (Tekin, 2023). Systematic review studies on AI in education are also increasing (e.g., Casal-Otero et al., 2023; Kavitha & Joshith, 2024; Tahiru, 2021; Zhai et al., 2021), indicating the field's growing maturity. With rapid advancements in AI, it is crucial for educators to understand how to effectively utilize these techniques to create optimal learning opportunities.

Although the consideration of AI in mathematics education research may seem new, it is important to remember that it has a history spanning approximately 60-70 years. One of the earliest programs in AI is Slagle's program, developed in 1961 as part of his doctoral dissertation to perform indefinite integration (Slagle, 1963). SAINT (Symbolic Automatic INTEgrator) demonstrated problem-solving abilities comparable to those of a freshmen calculus student. It successfully solved 84 out of 86 problems, 54 of which were drawn from Massachusetts Institute of Technology final exam questions. Although SAINT is considered an early AI system, it is designed for a limited scope to find antiderivatives of various functions. Today, however, AI technology has

Address of Corresponding Author

Yasin Memiş, PhD, Niğde Provincial Directorate of National Education, 51200, Niğde, Türkiye.

✉ ysnmemis@gmail.com

How to cite: Memiş, Y. (2025). Examining the potential and pitfalls of AI in problem solving. *Journal of Pedagogical Research*, 9(2), 177-190. <https://doi.org/10.33902/JPR.202532099>

reached a much more advanced level, performing a wide range of tasks, from language translation and graphic generation to programming and solving various mathematical problems. On the other hand, this multi-functionality can also affect performance, particularly in skills such as problem-solving (Plevris et al., 2023). ChatGPT, which was made available by OpenAI in November 2022, had a great impact in different fields in general and brought AI to the forefront again in many fields, including mathematics education. The ability of AI tools to provide information across different domains has paved the way for new research aimed at evaluating their performance.

AI tools such as ChatGPT have shown promising potential in enhancing mathematical education, offering both benefits and challenges that merit critical evaluation (Rane, 2023). The integration of these tools into educational settings can positively transform students' approaches to mathematical problem-solving by offering interactive learning experiences, real-time feedback, and personalized educational content (Wardat et al., 2023). ChatGPT, for instance, excels in generating contextually relevant responses, guiding students through complex problem-solving processes, and adapting to individual learning styles. This capability not only supports conceptual understanding and problem-solving skills but also fosters a more personalized and effective educational experience (Dao & Le, 2023).

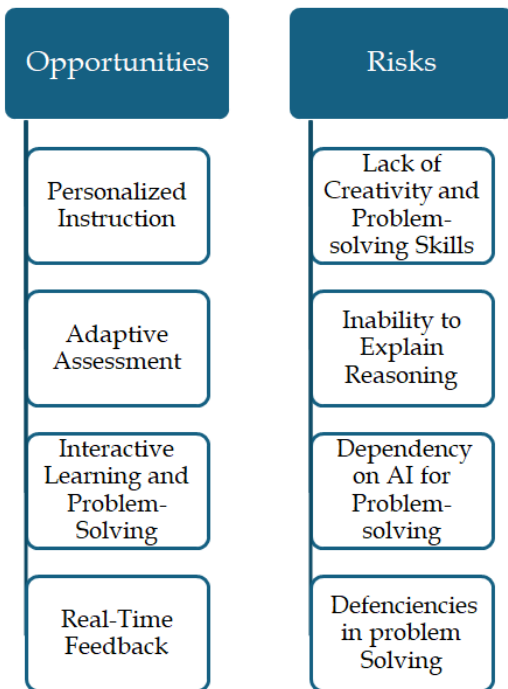
However, the deployment of AI in education is not devoid of limitations and requires careful consideration to mitigate potential risks. One significant concern is the accuracy and reliability of AI-generated solutions. Inaccuracies can mislead students, potentially reinforcing misconceptions and impeding learning progress (Dwivedi et al., 2023). Moreover, an overreliance on AI for problem-solving could diminish critical thinking and independent problem-solving skills among students, leading to superficial understanding rather than deep mathematical insights (Sánchez-Ruiz et al., 2023). Ethical concerns also arise, including issues related to data privacy, algorithmic bias, and the equitable use of technology across diverse student populations (Sallam, 2023). Balancing the innovative capabilities of AI tools with these challenges is crucial to harnessing their full potential in education without compromising educational integrity or equity.

1.1. The Impact of AI on Mathematic Education: Opportunities and Risks

With the advent of new technologies in mathematics education, we continue to grapple with familiar questions: How do we successfully integrate calculators into teaching? How do we effectively use the internet in mathematics education? Now, as mathematics educators, we must also develop a pedagogy that is specifically oriented towards AI (Engelbrecht & Borba, 2024).

The integration of AI into education, particularly in the field of mathematics, must be approached with a balanced consideration of both its opportunities and risks, as with any technology. It has significant transformative potential in mathematics learning; it can offer personalized instruction, adaptive assessments, and real-time feedback tailored to individual student needs and learning styles (Wardat et al., 2023). These advancements could significantly enhance the learning experience, making it more engaging and effective. However, the integration of AI into mathematics education also brings some challenges. Over-reliance on AI may lead to a decline in students' problem-solving skills and creativity, as they may become dependent on automated solutions rather than developing their own (Opesemowo & Ndlovu, 2024). Moreover, the limitations of AI, particularly in problem-solving, including its potential for making errors, overlooking conceptual contexts, and focusing directly on outcomes rather than reasoning, should not be ignored. Therefore, a cautious approach is essential to harness AI's power while minimizing potential risks, ensuring it serves as a stronger tool in mathematics education. The opportunities and risks associated (National Council of Teachers of Mathematics [NCTM], 2024; Opesemowo & Ndlovu, 2024) with AI in math education are as summarized in Figure 1.

Figure 1
Opportunities and Risks of AI in Mathematics Education



In light of recent technological advancements, the use of generative AI (GenAI) has become increasingly significant in mathematics education (Baidoo-Anu & Owusu, 2023; Plevris et al., 2023). GenAI refers to AI systems that can create new content, including text, images, code, and mathematical solutions, based on patterns learned from large datasets (Miao & Holmes, 2023). GenAI tools, such as ChatGPT, supported by their continuously updated databases, can also provide learning opportunities in mathematics education, particularly within the problem-solving process. For example, ChatGPT solved 60 math questions from the Taiwan's educational exams with an approximate accuracy rate of 90% (Li et al., 2023). The Comprehensive Assessment Program for Junior High School Students (CAP) categorizes mathematical abilities into three levels: A (proficient), B (basic), and C (needs improvement), with each level further subdivided into three sub-levels (B, B+, B++). Given that the average accuracy rate is 90%, which falls within the A+ range (88-94 points), it can be stated that ChatGPT's performance is at the proficient level (Li et al., 2023). Furthermore, expert math teachers who reviewed the answers pointed out that ChatGPT could provide more detailed and step-by-step solutions than those found in textbooks. Such studies indicate that AI tools hold promising potential for the future of mathematics education.

However, alongside the opportunities this technology offers, certain risks must also be acknowledged. As illustrated in Figure 1, AI's role in mathematics education, especially in problem-solving, requires careful consideration of its limitations. Even when AI tools correctly process inputs in mathematical problem-solving, they might struggle to provide reliable and accurate solutions, even for simple problems (Guler et al., 2024; Lewkowycz et al., 2022; Plevris et al., 2023). Frieder et al. (2023) have demonstrated that ChatGPT's performance in university-level mathematics can be inferior to that of typical students. They have particularly highlighted the necessity for enhancements in areas such as proof generation and high-quality computational outputs. Therefore, it is crucial to always verify the accuracy and reliability of AI-generated outputs. This research will present an analysis of the types of errors AI tools can make in mathematical problem-solving and how these errors occur.

Studies on the problem-solving ability of AI indicate that its performance varies depending on the AI tool used, its level of training, and the type of questions asked (Frieder et al., 2023). Research on AI performance in different areas of mathematics generally shows higher success in algebra,

numbers, and procedural tasks, while performance decreases in reasoning-intensive topics like geometry or in complex problem-solving scenarios (Wei, 2024). Skills such as interpreting graphs and diagrams or visualizing spatial relationships can challenge the performance of text-based AI tools (Wei, 2024). Moreover, in questions requiring complex reasoning and cognitive processes, if the AI system's background (the data it has been trained on) is not adequately developed or lacks sufficient data, its problem-solving performance will be affected. Sinha et al. (2024) demonstrated this with AlphaGeometry, a trained neuro-symbolic model, which successfully solved 25 out of 30 problems from the International Mathematics Olympiad. This result highlights that focused training on specific mathematical topics can enhance the problem-solving capabilities of AI systems.

AI tools demonstrate varying problem-solving abilities depending on the topic and question types. Therefore, this study focuses on proportional reasoning, which has not been previously addressed in AI research. Proportional reasoning is a critical skill in mathematics curricula in most countries (Memiş, 2022) and serves as a fundamental step for understanding advanced mathematical concepts such as algebra, geometry, statistics, and probability (Vanluydt et al., 2022). Lesh et al. (1988) describe proportional reasoning as a cornerstone for children's basic arithmetic skills and their progression to higher-level mathematical understanding. However, studies reveal that students frequently make conceptual errors in this area, making it a common challenge in mathematics education (Lamon, 2020). Examining the effectiveness of AI-based tools in solving proportional reasoning problems can help improve their role in mathematics education. Understanding the resources, knowledge, and problem-solving strategies (heuristics) a problem solver employs provides valuable insights into their approach to problem-solving (Schoenfeld, 2011). Measuring proportional reasoning skills typically involves analysing solutions to problems requiring various skills, such as distinguishing between additive and multiplicative relationships or applying qualitative reasoning (Cramer et al., 1993; Lamon, 2007). Therefore, examining how AI tools solve these problems can be used both to develop deep learning processes in mathematics education and to evaluate the pedagogical potential of AI. Furthermore, teachers' awareness of the potential of AI tools is essential for effectively integrating these technologies into classroom practices (Mohamed et al., 2022). Particularly, the extent to which AI tools frequently used by students for one-on-one interaction provide accurate and precise information and to what degree they can address conceptual misunderstandings remains a critical area of investigation.

This research examines the effectiveness of tools like ChatGPT-4, Gemini, and CoPilot in mathematics education, particularly regarding their ability to solve problems in areas where students commonly struggle. Specifically, the research focuses on the following question:

RQ) To what extent can ChatGPT-4, Gemini and CoPilot solve questions where students frequently make errors in proportional reasoning?

2. Method

The study utilised questions that students commonly erred on, including conceptual inaccuracies. 14 questions (see Appendix for all questions) requiring proportional reasoning skills were selected and solved using the ChatGPT-4, Gemini, and CoPilot. These tools are among the most widely utilized globally (Mardikar, 2024) and are freely accessible. Additionally, these tools have been widely utilised in related studies to examine problem-solving processes (Ergene & Ergene, 2024; Oh et al., 2024) and provide detailed solutions to problems presented in both textual and visual formats.

Each question was posed directly to these chatbots in separate conversation tabs, allowing for the recording of the first responses provided by these tools. In the initial stage, chatbot responses were assessed in two categories: correct and incorrect. A response was deemed correct only if both the outcome and the reasoning were mathematically accurate; otherwise, it was classified as incorrect. In the subsequent stage, incorrect responses were analysed to determine the sources of errors.

Additionally, no further guidance or contextual information was given to the bots during this process. This method was chosen to minimize the impact on the problem-solving approach of AI tools and to assess their capabilities in the simplest manner. Lastly, it should be considered that the answers given by these tools might change in subsequent attempts due to their continuous updates. To examine problem-solving processes, ChatGPT-4, Gemini, and Copilot were utilised in this study, with all tests conducted in September 2024. The analysis and evaluation of their impact on problem-solving strategies are based on the tests performed during this period. It should be noted that since then, there may have been developments and improvements in the responses provided by chatbots to these questions.

3. Findings

The answers obtained from these tools were evaluated within the scope of proportional thinking skills, and it was examined whether the mistakes made by the students were also made by AI tools. Thus, the performances of these three tools could be evaluated. Table 1 shows the accuracy of the answers given by these three tools to 14 questions.

Table 1
AI Tools Performance on 14 proportional reasoning questions

Questions	ChatGpt-4	Gemini	CoPilot
1	+	+	+
2	+	+	+
3	-	-	-
4	-	-	-
5	-	-	-
6	+	+	-
7	+	+	+
8	+	-	-
9	+	+	+
10	+	+	+
11	+	-	-
12	+	+	+
13	-	-	-
14	+	+	+
Accuracy rate	71%	57%	50%

Upon evaluating the table, it can be stated that ChatGPT-4 generally performed better compared to the other tools. Furthermore, it was observed that ChatGPT-4 provided more explanatory and step-by-step solutions than the other tools. In Figure 2, the answers given by three AI tools regarding question 6 are shown in Figure 2.


When analysing the responses in Figure 2, it is evident that CoPilot provided an incorrect answer compared to the other tools. Additionally, ChatGPT-4's response is observed to be more explanatory than other AI tools. In particular, it more clearly explains that an increase in the number of people does not affect the music play speed.

Another finding regarding the performance of AI tools is related to the extent to which they provide explanatory answers. The word count distribution among three AI tools—ChatGPT-4, Gemini, and Copilot—in response to 14 questions reveals that ChatGPT-4 stands out with a word count of 1401, compared to 880 for Gemini and 1069 for Copilot. This difference can be interpreted as an indication that ChatGPT-4 offers more detailed and explanatory answers than the other tools.


Figure 2


Comparison of ChatGPT-4, Gemini, and Copilot Responses to Question 6 (Presented in Sequential Order)


A group of five people plays a piece in 10 minutes. How many minutes would a group of ten people take to play the same song?

 To answer this question, it's important to understand whether the task of playing a piece is affected by the number of people performing it. In the context of a musical performance, the duration of the piece is typically fixed and does not depend on the number of musicians, assuming the composition and arrangement allow for that number of musicians to play together without changing the fundamental structure or duration of the piece.

If the piece originally takes 10 minutes to play, it will still take 10 minutes to play with ten people, assuming all are contributing to a coordinated, simultaneous performance. Therefore, if a group of ten people plays the same song, it would also take them 10 minutes, just like it did for the group of five.


 A group of five people plays a piece in 10 minutes. How many minutes would a group of ten people take to play the same song?

[Taslakları göster](#) 


 **The number of people playing doesn't affect the length of the song.**

If a group of 5 people can play a piece in 10 minutes, then a group of 10 people would also take **10 minutes** to play the same song.

The number of musicians doesn't change the duration of the piece itself.

 **You**

A group of five people plays a piece in 10 minutes. How many minutes would a group of ten people take to play the same song?

 **Copilot**

If a group of **five** people can play the piece in **10 minutes**, then the rate at which they play is

$$\frac{1}{10}$$

of the piece per minute.

Now, let's consider a group of **ten** people. Since the number of people has doubled, the rate at which they play will also double. Therefore, the group of ten people will play the same song in **5 minutes**. 🎵🎶

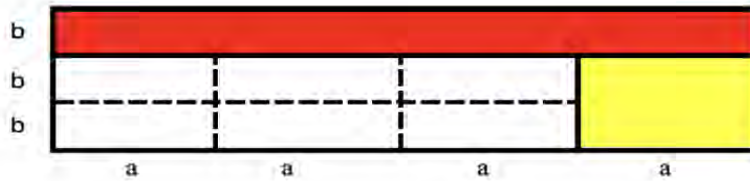
3.1. Evaluating Error Types

In this section, an evaluation of the incorrect answers provided by AI tools has been conducted. As a result of this evaluation, it was found that AI tools can make various types of errors. A classification of these errors is presented at the end of the section.

AI tools gave incorrect answers to these questions because they could misinterpret visual inputs while solving questions with visual content. For example, in the question in Figure 3, which involves an area ratio, ChatGPT-4 failed to correctly identify and calculate the areas, leading to an incorrect ratio.

Figure 3

ChatGpt-4's wrong response to the Visual-Based Question *(Question 3)



What is the ratio of the red area to the yellow area?"

To find the ratio of the red area to the yellow area in the diagram, we need to identify the dimensions and areas of each colored section:

1. Red Area Calculation:

- The entire top rectangle spans 4 units of "a" horizontally (as each section marked "a" horizontally spans the entire width of the individual sections underneath).
- The height of the red rectangle is "b".
- Thus, the area of the red rectangle is $4a \times b = 4ab$.

2. Yellow Area Calculation:

- The yellow section spans 1 unit of "a" horizontally.
- The height of the yellow rectangle is also "b".
- Thus, the area of the yellow rectangle is $a \times b = ab$.

Ratio of Red Area to Yellow Area:

- The ratio is $\frac{\text{Red Area}}{\text{Yellow Area}} = \frac{4ab}{ab}$.
- Simplifying this gives a ratio of 4:1.

Therefore, the ratio of the red area to the yellow area is 4:1.

When examining the solution in Figure 3, it is noted that the visual was correctly divided into four equal parts horizontally, and the red area was accurately calculated. However, while the division "a" was correctly used for the yellow area horizontally, the height of "2b" was not

recognized. This problem is probably due to the fact that the yellow area is not clearly segmented by dashed lines, so that the AI cannot clearly identify it. Similarly, it was observed that the other tools could not correctly include the lengths of the coloured segments in the solution. This shows that these tools still make mistakes, especially in visual representations, and therefore cannot reach the correct result.

Upon examining the responses, another type of error observed is incorrect reasoning. In particular, correctly identifying the relationship between quantities is crucial in proportional reasoning. In other words, determining how the quantities in a problem are related (additively, multiplicatively, or not at all) is a critical step in solving the proportional reasoning problems. Indeed, AI tools have been showed to same error. An example of this is shown in Figure 4.

Figure 4

*Incorrect Response by ChatGPT-4 *(Question 5)*

Sue and Jack were running at the same speed around a track. Jack started first.
When Jack had run 4 laps, Sue had run 2 laps. When Sue had completed 6 laps, how many laps had Jack?

🌀 To find out how many laps Jack had run when Sue completed 6 laps, let's analyze the situation.

From the problem, we know:

- When Jack had run 4 laps, Sue had run 2 laps.

This tells us that Jack runs twice as many laps as Sue in the same amount of time because when Jack ran 4 laps, Sue only ran 2 laps. Therefore, their lap ratio is 2:1 in favor of Jack.

Now, we need to find how many laps Jack runs when Sue runs 6 laps. Since Jack's speed in laps is twice that of Sue's, we can apply the ratio directly:

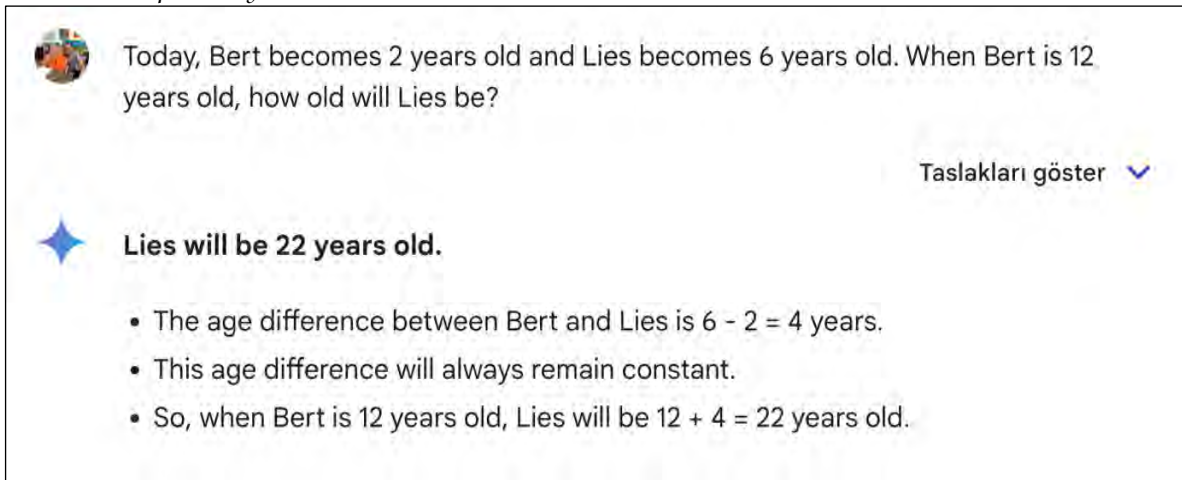
- If Sue runs 6 laps, then Jack runs $6 \text{ laps} \times 2 = 12 \text{ laps}$.

Therefore, when Sue had completed 6 laps, Jack had completed **12 laps**.

Upon examining the problem depicted in Figure 4, it is evident that the question demonstrates an additive relationship rather than a multiplicative one. In other words, there is a 2-lap difference between Jack and Sue, who maintain the same speed. Thus, when Sue completes his 6th lap, Jack will finish her 8th lap (6+2). However, ChatGPT-4 and other AI tools incorrectly interpreted this as a multiplicative relationship, assuming the other runner was twice as fast, leading to an erroneous response. This error is related to the inadequate distinction between additive and multiplicative relations, a challenge frequently discussed in the mathematics education as a critical issue in students' understanding of proportional reasoning.

Another error is illustrated in Figure 5. This incorrect solution is reported because it contains the only arithmetic error found across all solutions analysed. Although Gemini consistently applied a systematic and correct logic to the problem, it ultimately made a computational error by calculating $12+4$ as 22, leading to an incorrect answer. This example should be seen as further evidence of the need for careful examination of AI-generated solutions.

Figure 5

Incorrect Response by Gemini


Today, Bert becomes 2 years old and Lies becomes 6 years old. When Bert is 12 years old, how old will Lies be?

Taslakları göster

Lies will be 22 years old.

- The age difference between Bert and Lies is $6 - 2 = 4$ years.
- This age difference will always remain constant.
- So, when Bert is 12 years old, Lies will be $12 + 4 = 22$ years old.

After reviewing all the errors, they have been categorized into three main types. The first category involves the incomplete evaluation of data, particularly in visual questions. The second pertains to the use of incomplete or incorrect reasoning when assessing the questions. Lastly, arithmetic errors were observed in only one incorrect response. These error types are illustrated in Table 2.

Table 2

Frequencies of Error Types

<i>Error Type</i>	<i>Frequency (%)</i>
Incomplete evaluation of data	9 (52%)
Incomplete and/or incorrect reasoning	7 (41%)
Arithmetic errors	1 (6%)

4. Discussion and Conclusion

In this research, the performance of three leading chatbot systems—ChatGPT-4, Gemini, and Copilot—in solving problems that require proportional reasoning was compared. The research utilized question types emphasized in the mathematics education literature as common sources of student errors. The objective of this study was to evaluate the effectiveness of these tools in solving proportional reasoning problems. The findings indicate that all three chatbots possess a certain level of ability to understand and process these types of problems, though there are some exceptions and limitations. It was observed that their capabilities in this domain have specific constraints, suggesting that the reliability of these tools in performing such tasks may not always be consistent. The analysis revealed that ChatGPT answered more questions correctly and provided more detailed explanations. However, the chatbots also produced incorrect answers due to errors in reasoning, inability to interpret visual data, or calculation mistakes.

Research on the mathematical problem-solving ability of AI demonstrates that performance varies depending on the AI tool used, its level of training, the topic, and the type of question (Frieder et al., 2023; Plevris et al., 2023). While language models have made significant progress in recent years, they still face challenges in processing graphical data, understanding complex mathematical concepts, and solving problems requiring reasoning (Dao & Le, 2023). Studies on AI performance across different areas of mathematics indicate higher success rates in algebra, numbers, and procedural tasks, whereas performance declines in geometry or reasoning-based complex problem-solving scenarios (Wei, 2024). It was observed that AI tools provided correct answers to proportional reasoning questions involving procedural processes and algorithm-based tasks. However, errors occurred when solving questions requiring the interpretation of data from tables or figures. This may be due to the limitations of text-based AI tools in interpreting visual

information (Wei, 2024). Although this study identified only a single arithmetic error, research has also reported similar mistakes. For example, Bubeck et al. (2023) highlighted that even in simple problems like $7 * 4 + 8 * 8$ ChatGPT-4 could provide an incorrect answer, such as 88. Researchers suggest that this issue arises from the AI's limited memory when dealing with these problems. They found that success rates increased significantly when questions were asked in a way that encouraged step-by-step reasoning, such as using the prompt "Let's think step by step to solve the expression, write out all intermediate steps, and only then produce the final solution". Another type of error encountered in the study arose in questions requiring reasoning skills. Such errors, resulting from incorrect or incomplete reasoning, have also been reported in other studies (Plevris et al., 2023). Since these tools are developed as language models, it should be acknowledged that they may exhibit limitations in logical or reasoning tasks. Therefore, future research could aim to address these limitations and improve the mathematical problem-solving capabilities of language models.

Moreover, as seen in examples, while the solutions generated by chatbots often appear detailed and professional, they can contain various errors upon closer examination. Among these errors, the hallucination effect—characterized by the generation of false or misleading information—is an inherent limitation of all language models and should be carefully considered (Janéafik & Dusek, 2024). The main issue with the hallucination effect is that AI language models can generate false information and present it in a persuasive and explanatory manner, making it appear accurate. This problem is also encountered in the process of solving maths problems (Lee, 2023; Pardos & Bhandari, 2024). Particularly among younger age groups or students newly learning a subject, such detailed and confident responses may be accepted as correct without critical evaluation. When interacting with chatbots, users must evaluate the reliability of these responses themselves. As these tools continually improve, their accuracy increases, which in turn raises user trust. This could exacerbate the problem of uncritical acceptance. It's important to remember that one of the fundamental steps in mathematical education is "verification," or in other words, answering the question "Why is this correct?" While reaching a result is a desired step in problem-solving, verifying a solution and generating creative solutions through different approaches also require a deep understanding, which are other crucial steps in the process.

In conclusion, while AI tools have achieved a certain level of success in solving mathematical problems, they may exhibit the same errors frequently made by students. While the capabilities of AI can often be impressive, like all technologies, it has its limitations. This study reveals that these technologies sometimes produce responses that may lack context or struggle to grasp the nuances of certain topics. Therefore, when evaluating the use of these technologies in education, it is crucial for field experts to highlight various limitations. By doing so, educators and other stakeholders can make more informed decisions regarding the integration of these technologies, developing strategies to maximize benefits while minimizing risks.

5. Limitations

This study is limited to a single mathematical reasoning skill using a small sample of questions. Therefore, to achieve more generalizable results, it is recommended to incorporate a larger set of questions and employ diverse research methodologies. Additionally, when evaluating the problem-solving capacity of AI tools, it is important to consider the inherent limitations associated with their probabilistic nature. Chatbots may generate different responses to the same problem each time it is posed. Moreover, factors such as the prompt method used and updated versions of the same AI tools can lead to variations in responses to identical questions.

Declaration of interest: The authors declare that no competing interests exist.

Ethical declaration: Since no human participants were involved in the study, ethics committee approval was not required.

Funding: The authors stated that they received no financial support for their study.

References

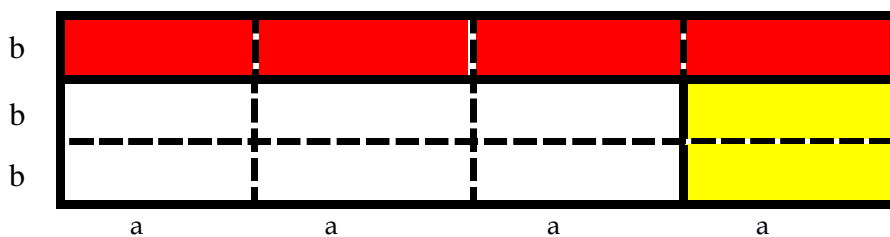
- Baidoo-Anu, D., Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of Chatgpt in promoting teaching and learning. *Journal of Artificial Intelligence*, 7(1), 52-62.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T. & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with Gpt-4* (arXiv:2303.12712). arXiv. <https://doi.org/10.48550/arXiv.2303.12712>
- Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K-12: A systematic literature review. *International Journal of STEM Education*, 10(1), 29.
- Cramer, K., & Post, T. (1993). Making connections: A case for proportionality. *The Arithmetic Teacher*, 40(6), 342-346.
- Dao, X. Q., & Le, N. B. (2023). *Investigating the effectiveness of ChatGPT in mathematical reasoning and problem solving: Evidence from the Vietnamese national high school graduation examination* (arXiv:2306.06331). arXiv. <https://arxiv.org/abs/2306.06331>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., & Wright, R. (2023). So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice, and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Engelbrecht, J., & Borba, M. C. (2024). Recent developments in using digital technology in mathematics education. *ZDM*, 56(2), 281-292.
- Ergene, O., & Ergene, B.C. (2024). AI ChatBots' solutions to mathematical problems in interactive e-textbooks: Affordances and constraints from the eyes of students and teachers. *Education and Information Technologies*, 30, 509-545. <https://doi.org/10.1007/s10639-024-13121-z>
- Frieder, S., Pinchetti, L., Griffiths, R.R., Salvatori, T., Lukasiewicz, T., Petersen, P.C., Chevalier, A., & Berner, J. (2023). *Mathematical capabilities of ChatGPT* (arXiv:2301.13867). arXiv. <https://doi.org/10.48550/arXiv.2301.13867>
- Guler, N., Dertli, Z. G., Boran, E., & Yildiz, B. (2024). An artificial intelligence application in mathematics education: Evaluating ChatGPT's academic achievement in a mathematics exam. *Pedagogical Research*, 9(2), 1-12. <https://doi.org/10.29333/pr/14145>
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(4), 523-545. <https://doi.org/10.1007/s13394-013-0083-6>
- Janéafik, A., & Dusek, O. (2024). The problem of AI hallucination and how to solve it. In F. Moriera (Ed.), *Proceedings of the 23rd European Conference on e-Learning* (Vol. 23, No. 1, pp. 122-128). Academic Conferences International. <https://doi.org/10.34190/ecel.23.1.2584>
- Kavitha, K., & Joshith, V. P. (2024). Pedagogical incorporation of artificial intelligence in K-12 science education: A decadal bibliometric mapping and systematic literature review (2013-2023). *Journal of Pedagogical Research*, 8(4), 437-465. <https://doi.org/10.33902/JPR.202429218>
- Lamon, S. J. (2007). Rational numbers and proportional reasoning: Toward a theoretical framework. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 629-668). Information Age Publishing.
- Lamon, S. J. (2020). *Teaching fractions and ratios for understanding: Essential content knowledge and instructional strategies*. Routledge. <https://doi.org/10.4324/9781003008057>
- Lee, M. A. (2023). Mathematical investigation of hallucination and creativity in GPT models. *Mathematics*, 11(10), 2320. <https://doi.org/10.3390/math11102320>
- Lesh, R., Post, T., & Behr, M. (1988). Proportional reasoning. In J. Hiebert, & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 93-118). National Council of Teachers of Mathematics.
- Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., (2022). Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35, 3843-3857. <https://doi.org/10.48550/arXiv.2206.14858>
- Li, P. H., Lee, H. Y., Cheng, Y. P., Starčić, A. I., & Huang, Y. M. (2023). Solving the self-regulated learning problem: Exploring the performance of ChatGPT in mathematics. In Y. M. Huang, & T. Rocha (Eds.),

- Innovative technologies and learning: Lecture notes in computer science* (pp. 77-86). Springer. https://doi.org/10.1007/978-3-031-40113-8_8
- Lobato, J., & Ellis, A. (2010). *Developing essential understanding of ratios, proportions, and proportional reasoning for teaching mathematics in grades 6-8*. National Council of Teachers of Mathematics.
- Mardikar, S., Aldhuhayyan, A., & Prabhu, V. V. (2024). Can the “art” of mathematical modeling in industrial engineering be automated by AI? In M. Rahman (Ed.), *Proceedings of the 9th North American Conference on Industrial Engineering and Operations Management* (pp. 1030-1041). IEOM Society International. <https://doi.org/10.46254/NA09.20240261>
- Memiş, Y. (2022). *Examination of secondary school textbooks from Turkey, Singapore, and Canada in the context of proportional thinking* (Publication no. 463446) [Master’s thesis, Anadolu University]. Council of Higher Education Thesis Center.
- Miao, F., & W. Holmes. (2023). *Guidance for generative AI in education and research*. UNESCO.
- Mohamed, M. Z. B., Hidayat, R., Suhaizi, N. N. B., Sabri, N. B. M., Mahmud, M. K. H. B., & Baharuddin, S. N. B. (2022). Artificial intelligence in mathematics education: A systematic literature review. *International Electronic Journal of Mathematics Education*, 17(3), 12132. <https://doi.org/10.29333/iejme/12132>
- National Council of Teachers of Mathematics. (2024). *Artificial intelligence and mathematics teaching*. Author.
- Oh, S., Yoon, J., Chung, Y., Cho, Y., Shim, H., & Kwon, O. N. (2024). Analysis of generative AI’s mathematical problem-solving performance: Focusing on ChatGPT 4, Claude 3 Opus, and Gemini Advanced. *The Mathematical Education*, 63(3), 549-571.
- Opesemowo, O. A. G., & Ndlovu, M. (2024). Artificial intelligence in mathematics education: The good, the bad, and the ugly. *Journal of Pedagogical Research*, 8(3), 333-346. <https://doi.org/10.33902/JPR.202426428>
- Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS ONE*, 19(5), e0304013. <https://doi.org/10.1371/journal.pone.0304013>
- Plevris, V., Papazafeiropoulos, G., Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: a comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *Artificial Intelligence*, 4(4), 949-969. <https://doi.org/10.3390/ai4040048>
- Rane, N. (2023). *Enhancing mathematical capabilities through ChatGPT and similar generative artificial intelligence: Roles and challenges in solving mathematical problems*. SSRN. <https://doi.org/10.2139/ssrn.4603237>
- Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11(6), 887. <https://doi.org/10.3390/healthcare11060887>
- Sánchez-Ruiz, L. M., Moll-López, S., Nuñez-Pérez, A., Moraño-Fernández, J. A., & Vega-Fleitas, E. (2023). ChatGPT challenges blended learning methodologies in engineering education: A case study in mathematics. *Applied Sciences*, 13(10), Article 6039. <https://doi.org/10.3390/app13106039>
- Schoenfeld, A. H. (2011). *How we think: A theory of goal-oriented decision making and its educational applications*. Routledge.
- Sinha, S., Prabhu, A., Kumaraguru, P., Bhat, S., & Bethge, M. (2024). *Wu's method can boost symbolic AI to rival silver medalists and alpha geometry to outperform gold medalists at IMO geometry* (arXiv:2404.06405). arXiv. <https://arxiv.org/abs/2404.06405>
- Slagle, J. R. (1963). A heuristic program that solves symbolic integration problems in freshman calculus. *Journal of the ACM*, 10(4), 507-520. <https://doi.org/10.1145/321186.321193>
- Stacey, K. (1989). Finding and using patterns in linear generalising problems. *Educational Studies in Mathematics*, 20(2), 147-164. <https://doi.org/10.1007/bf00579460>
- Tahiru, F. (2021). AI in education: A systematic literature review. *Journal of Cases on Information Technology*, 23(1), 1-20.
- Tekin, N. (2023). Artificial intelligence in education: a content analysis on trends in research from Türkiye. *Journal of Necmettin Erbakan University Ereğli Faculty of Education*, 5, 387-411.
- Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., & Verschaffel, L. (2005). Not everything is proportional: Effects of age and problem type on propensities for overgeneralization. *Cognition and Instruction*, 23(1), 57-86. https://doi.org/10.1207/s1532690xci2301_3
- Vanluydt, E., Verschaffel, L., & Van Dooren, W. (2022). The role of relational preference in early proportional reasoning. *Learning and Individual Differences*, 93, 102108. <https://doi.org/10.1016/j.lindif.2021.102108>
- Wardat, Y., Tashtoush, M. A., Ali, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), 13272. <https://doi.org/10.29333/ejmste/13272>

- Wei, X. (2024). Evaluating chatGPT-4 and chatGPT-4o: performance insights from NAEP mathematics problem solving. *Frontiers in Education*, 9, 1452570. <https://doi.org/10.3389/feduc.2024.1452570>
- Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J. B., Yuan, J., & Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity*, 2021, 8812542. <https://doi.org/10.1155/2021/8812542>

Appendix. The proportional reasoning questions

- 1) Yesterday you shared some cookies with some friends. Today, you share fewer cookies with more friends. Will everyone get more, less, or the same amount as they received yesterday? (Lamon, 2020, p.69)
- 2) Yesterday you shared some cookies with some friends. Today, you share more cookies with more friends. Will everyone get more, less, or the same amount as they received yesterday? (Lamon, 2020, p.69)
- 3) What is the ratio of the red area to the yellow area? (Lamon, 2020, p.114)



- 4) In the picture, a ladder with 2 steps is made using 8 matchsticks. If we want to make a ladder with 6 steps, how many matchsticks will we need? (Stacey, 1989 p. 148)



- 5) Sue and Jack were running at the same speed around a track. Jack started first. When Jack had run 4 laps, Sue had run 2 laps. When Sue had completed 6 laps, how many laps had Jack? (Hilton et al., 2013, p.539)
- 6) A group of five people plays a piece in 10 minutes. How many minutes would a group of ten people take to play the same song? (Hilton et al., 2013, p.542)
- 7) The locomotive of a train is 12m long. If there are 4 carriages connected to the locomotive, the train is 52m long. How long would the train be if there were 8 carriages connected to the same locomotive? (Van Dooren, 2005, p.77)
- 8) Today, Bert becomes 2 years old and Lies becomes 6 years old. When Bert is 12 years old, how old will Lies be? (Van Dooren, 2005, p.65)
- 9) Three cups contain different amounts of water and sugar. Cup A is full of water with 3 lumps of sugar. Cup B is half full of water with 2 lumps of sugar. Cup C is one-third full of water with 1 lump of sugar. After the lumps of sugar are stirred in, which cup will be the sweetest? (Hilton et al., 2013, p.543)
- 10) Find X in the proportion $x:4=20:35$

Appendix continued

11) **Clown's Journey Problem**

The following table represents the clown's location from home, and his time:

Location (cm)	Elapsed Time (sec)
-5	2
20	12
40	20
52.5	25
140	60

a. Did the clown walk at the same speed, or did he speed up or slow down? How can you tell? (Lobato & Ellis, 2010 p.76)

12) y is directly proportional to x^3 . When $x = 2$, $y = 24$. What is the value of x when $y = 192$

13) Which detergent is more cost-effective to choose? (Hilton et al., 2013, p.541)

Washing Powder A	Washing Powder B
Price: \$4.00	Price: \$6.50
Quantity: 1 kilogram	Quantity: 1.5 kilograms
Number of washes: 20 washes	Number of washes: 30 washes

14) If a is directly proportional to b , and b is inversely proportional to c , what is the relationship between a and c ?