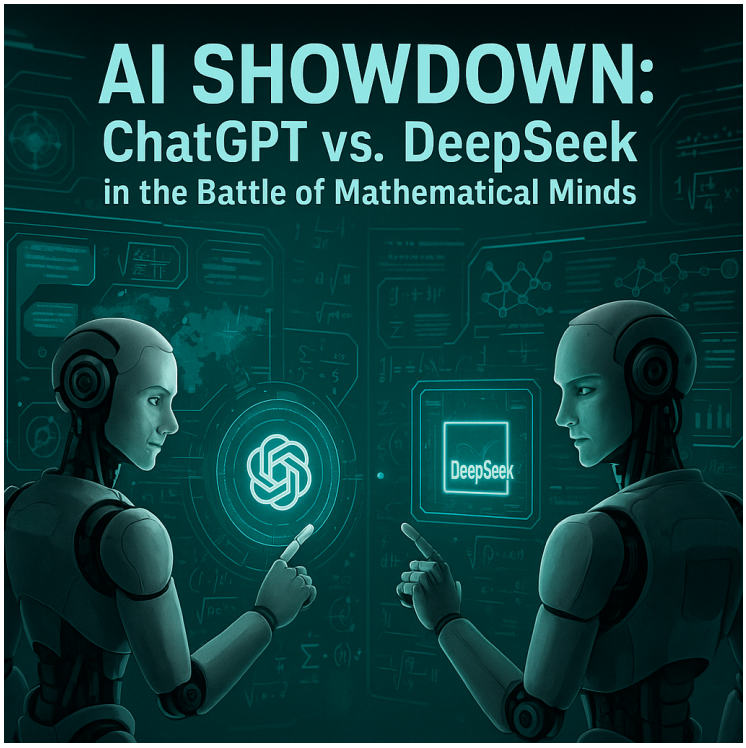
AI Showdown: ChatGPT vs. DeepSeek in the Battle of Mathematical Minds

American University of Beirut
English 206
Section 4
Group 5

To Dr. May Mikati
April 19, 2025

Sara Malaeb
Miguel Kattouah
Ahmad Jouni
Karim Assi



Abstract

In today's era of rapid technological progress, artificial intelligence tools have become global. Among their many applications, using AI to handle complex mathematical problems stands out as particularly promising. With a growing number of competing models, each offering unique capabilities and specialized functions, selecting the most suitable tool for a specific task can be challenging. Both ChatGPT and DeepSeek are strong competitors in this space, demonstrating impressive mathematical problem-solving skills, yet each has its strengths and limitations depending on the field and context in which they are applied. The purpose of this research is to compare the feasibility of ChatGPT and DeepSeek in solving complex mathematical problems, evaluating their correctness, consistency, efficiency, reliability, and cost of maintenance. We begin by introducing the growing role of artificial intelligence in mathematical problem-solving and provide an overview of both models, including their features and intended use cases. The methods employed in this research include direct testing with advanced mathematical problems across various fields: calculus and algebra. The central part of the report consists of a detailed analysis of each model's capability to solve problems accurately, explain solutions, respond efficiently, and maintain consistent availability. We also assess the financial implications of using these chatbots in academic and professional settings. The final section of the report presents a comparative summary and provides practical recommendations on when and how each model should be applied. Our findings suggest that while both chatbots are capable of accurately solving complex mathematical problems, their effectiveness and suitability vary depending on the user's preferences, with ChatGPT excelling in efficiency and clarity, and DeepSeek offering more detailed, educational-style explanations.

Table of Contents

List of Illustrations

Table of Figures

Table of Tables

I. Introduction

Historically, mathematical problem-solving has been a difficult and prolonged endeavor. Collaboration among math experts typically occurred through traditional postal communication, significantly delaying progress and solutions to algebraic, geometric, and various other mathematical questions. Today, artificial intelligence (AI) has dramatically transformed this landscape, providing solutions within mere seconds and drastically accelerating mathematical advancements. Modern AI systems, especially large language models (LLMs) like ChatGPT and DeepSeek, utilize immense databases, powerful computational servers, and efficient algorithms to rapidly and effectively assist users in solving complex mathematical challenges. However, despite the undeniable advantages in speed and efficiency, the reliance on AI systems also raises concerns regarding the accuracy, consistency, and possible biases inherent in AI-generated outputs (Katta, 2025). Moreover, AI's role has influenced various sectors, including education, engineering, and regulatory sciences. According to Baran and Tong (2024), establishing clear governance frameworks and ensuring transparency are crucial, especially considering the diverse degrees of autonomy exhibited by different AI models. The increased use of AI models like ChatGPT and DeepSeek in educational settings further emphasizes the need for thorough evaluations regarding their capabilities and limitations (Ayele, 2024).

Our feasibility study examines and compares the effectiveness of ChatGPT and DeepSeek in solving advanced mathematical problems. We assess the LLMs on exercises from simple algebra problems to complex calculus and algebra questions. The report will examine ChatGPT and DeepSeek's efficiency in solving problems accurately. Furthermore, we consistently assess the LLMs' reliability in solving math problems. Finally, we evaluate the cost of maintaining such LLMs. The goal is to help students, teachers, and professionals in STEM decide which model fits their needs best, based on practical criteria.

II. Methods

A. Information Gathering

1. Primary Sources

We started our research by asking the different LLMs to answer the following 3 mathematical questions. The first is an algebra question in the grade 9 Lebanese "Ahlia" book that depicts a system of 2 equations. To be able to answer the question, the LLMs should use logical thinking to identify the main problem, solve for "g" and "b" in each of the two cases, and then use human-like reasoning to identify which one of the girls is lying.

Question 1:

Sabine: "In my class, there are 28 students. If I multiply the number of boys by 3 and the number of girls by 2, and if I add these two results, I obtain 60."

Carole: "In my class, there are 30 students. If I multiply the number of boys by 2 and the number of girls by 5, and if I add these two results, I obtain 95."

One of the girls is not telling the truth. Which one? Why?

Referring to Figures 3 and 4 in the appendix, both LLMs correctly solved the problem and identified that the number of students should be whole. Similarly, they were both able to deduce that Carole is the one who is lying. However, as seen by their responses, ChatGPT's response is more straightforward, as it did not explain the answer or provide detailed steps. On the other hand, DeepSeek explained its reasoning and verified its answer on multiple fronts, ultimately reaching a clear answer. As this is a simple problem, our analysis would not be based on correctness or consistency but rather on the efficiency of such an AI and how reliable it would be in solving logical reasoning problems.

On the other hand, we assessed the problems on a complex mathematical problem found in our Calculus Course. The problem is taken from a Math 202 Calculus Book.

Question 2: Find the area of the surface cut from the paraboloid $z= x^2 + y^2$ by the plane $z= 2$.

As seen in Figures 5 and 6 in the appendix, ChatGPT and DeepSeek solved the problems differently. ChatGPT used the surface area formula, found the partial derivatives with respect to x and y, and then integrated after replacing the partials with 2x and 2y, respectively. Afterwards, ChatGPT changed to polar coordinates and integrated the function over the disk region. The model solved the integral through a substitution and then got the final area to be square units. However, DeepSeek used a different approach closer to the method we learned in class. DeepSeek automatically changed to polar coordinates and computed the partial derivatives. In this problem, DeepSeek found the cross product of the partial derivatives of the polar coordinates with respect to r. It found the magnitude and then computed the surface area using the same integral as ChatGPT. It then followed the same substitution and reached the same answer.

Therefore, ChatGPT used a simpler approach than DeepSeek; the former assumed that the person solving the exercise knew the surface area formula. On the other hand, DeepSeek completely avoided the formula and used a different approach, which included more complex math with basic concepts. In terms of explanation, ChatGPT explained the use of its method better than DeepSeek. Although ChatGPT assumed we knew all concepts, it explained the purpose and use of all the formulas employed. On the other hand, DeepSeek moved into the second phase of the solution (after part 2) without explaining the reasoning behind its steps.

2. Secondary Sources (Reports and Benchmarking Studies)

We examined two separate peer-reviewed academic papers and benchmarked them, gaining a comprehensive understanding of ChatGPT and DeepSeek's capabilities in mathematical problem-solving. These studies gave benchmark evaluation and insights into what went into each model's internal training methodologies.

- ChatGPT: Mathematical Capabilities and Evaluation

ChatGPT and Deepseek in solving mathematical problems

This report introduces two carefully constructed mathematical benchmarks used to assess ChatGPT's ability to reason in mathematical language, covering topics from undergraduate classes to graduate-level questions.

- Strengths of ChatGPT:
  - Excels at definitions, factual questions, and undergraduate math.
  - Useful as a mathematical assistant
- Limitations:
  - Struggles with symbolic integration, proof completions, and graduate-level content.
  - Performs inconsistently on complex tasks that require a deep mathematical insight.
- Human Experts Rating: Advanced problem sets — average rating of ~3.2/5.

ChatGPT's mode of thinking and decision-making is based on a "black-box". This phenomenon is defined as the ability to decide based on previous experience and problem-solving data. Although this seems like a very human and reasonable mode of thinking, thus improving the deep-learning system, the inability to trace the system's thought process and see why it made this decision makes it feasibly impossible to get an accurate result at every turn. (Blouin, 2023) Moreover, this also depends on the number of situations the learning model has been exposed to. Theoretically, if the AI model is exposed to every situation possible, we would get 100% accurate results. Realistically, this is not possible.

- DeepSeekMath: A Math-Focused AI Model:

This paper introduces DeepSeekMath 7B, a model designed for math problem-solving. A staggering 120 billion math-focused examples from the internet were used to train it, and it was then improved using a deft fine-tuning method called Group Relative Policy Optimization (GRPO).

- Highlights:
  - On the MATH benchmark, scored 51.7%

- 60.9% consistency, and produced stable, reliable answers across multiple prompts.
- Good to use for math problems in different languages

- Notes:
  - Built on a model trained with a specific code, which helps it track reasoning and correct old mistakes, and reduce error rates
  - The model is based on GRPO, a more efficient and less computationally intensive method for improving reasoning.

The developers of DeepSeekMath emphasize that the model was trained on over 120 billion math-related problems, giving it a strong foundation in problem-solving across a range of mathematical fields. Unlike more general language models, DeepSeekMath was optimized using a training technique (GRPO), which enhances its reasoning capabilities by refining responses across multiple attempts. This method improves consistency and makes the model more efficient. (Shao et al.,2024)

B. Assessment Criteria:

1. Correctness & Efficiency:

We investigated the models' responses based on two criteria:

-Correctness: Did the AI solve the math problem correctly?

-Efficiency: Did it solve without any interruptions or confusing reasons?

For our primary sources:

In Question 1, both ChatGPT and DeepSeek computed the correct answer; however, the logical structure of each model differed, where ChatGPT provided a brief response, while DeepSeek solved this problem using a step-by-step approach. Also, ChatGPT was more effective in simplicity and speed; DeepSeek matched teaching methods, which are more helpful for students.

In Question 2, both AI models solved it with no mistakes, but the method used for solving this problem was distinct.

From secondary sources:

On the competition-level math benchmark, DeepSeek scored a 51.7% accuracy, surpassing ChatGPT (42–45%). Additionally, DeepSeek was better than ChatGPT in structured step-by-step problems.

In Summary, both models were accurate and effective. DeepSeek showed more steps, while ChatGPT was faster, assuming prior knowledge (you can ask ChatGPT to show the steps, but it assumes that the user needs the answer only)

2. Consistency & Reliability:

Based on our testing and academic studies, ChatGPT and DeepSeek provided the same answers. With a self-consistency score of 60.9%, DeepSeek consistently generated the right results over several runs and prompt variations. “DeepSeek has drawn global attention not just for its performance, but for its accessibility—offering advanced reasoning capabilities in math and science while remaining open and cost-effective, a combination that has thrilled researchers across disciplines.” (Gibney, E. (2025), Nature).

On more difficult problems, ChatGPT was less reliable; occasionally, when we changed the prompt, the answer changed, which caused some confusion among users. Although ChatGPT shows excellence at providing quick answers, this could sometimes lead to gaps in logic. In the end, DeepSeek gives more consistent solutions, and ChatGPT is reliable for basic questions but can change answers if the wording is creative or tricky. A recent study in Popular Science revealed that ChatGPT's performance on math problems declined over time, with accuracy rates dropping significantly between versions. Researchers found that while earlier iterations of ChatGPT-4 were more reliable in providing correct solutions, later

updates sometimes produced incorrect or inconsistent answers, especially for multi-step problems. (Paul, 2023).

3. Cost of Maintenance:

According to the studies, the most advanced version of ChatGPT requires a paid subscription and is a closed source (you cannot modify its software); DeepSeek is significantly less expensive to run.

In terms of training, ChatGPT utilizes a large dataset to learn, then gets extra training where humans review the answers and guide it to better solutions. This could make ChatGPT smarter and more aligned with human thinking. On the other hand, DeepSeek uses a lighter method (GRPO), which evaluates the AI's responses by comparing them within a group of responses. This technique requires less memory and makes DeepSeek cost-effective.

III. Results

Our analysis of ChatGPT and Deepseek was done using primary mathematical problems: an algebraic problem from the Lebanese Grade 9 "Ahlia" textbook, and a Calculus II surface area problem containing a paraboloid. The performance of these chatbots was assessed using three evaluation dimensions: Correctness and Efficiency, Consistency and Reliability, and Cost of Maintenance.

A. Correctness and Efficiency

1. Performance Metrics in Different Categories

- Algebraic Problem: Both chatbots derived mathematically correct answers. They identified that the number of students should be positive integers (whole numbers), and both concluded that Carole must be lying

according to logical reasoning from the system of equations. However, the nature of their solutions was different:

  - ChatGPT provided a brief answer, directly identifying the lie without too much elaboration. “ChatGPT’s strength lies in its ability to deliver concise, well-structured explanations that adapt to user intent, making it a powerful tool for quick learning and consistent application across varied scenarios.” (Lee, H. (2024), Anatomical Sciences Education)
  - DeepSeek provided a more layered response, in which it showed a step-by-step calculation and verified the logic from different perspectives.

- Calculus II Problem: Both ChatGPT and DeepSeek correctly solved the surface area problem. However, their methods of solving differed significantly:
  - ChatGPT used the surface area formula: $A=\iint 1+(zx)2+(zy)2dxdy$ and calculated partial derivatives. It then converted to polar coordinates to evaluate the integral.
  - DeepSeek utilized a vector-based geometric approach, converting the paraboloid into polar coordinates from the start, and computing the cross product of the partial derivatives to reach the final solution.

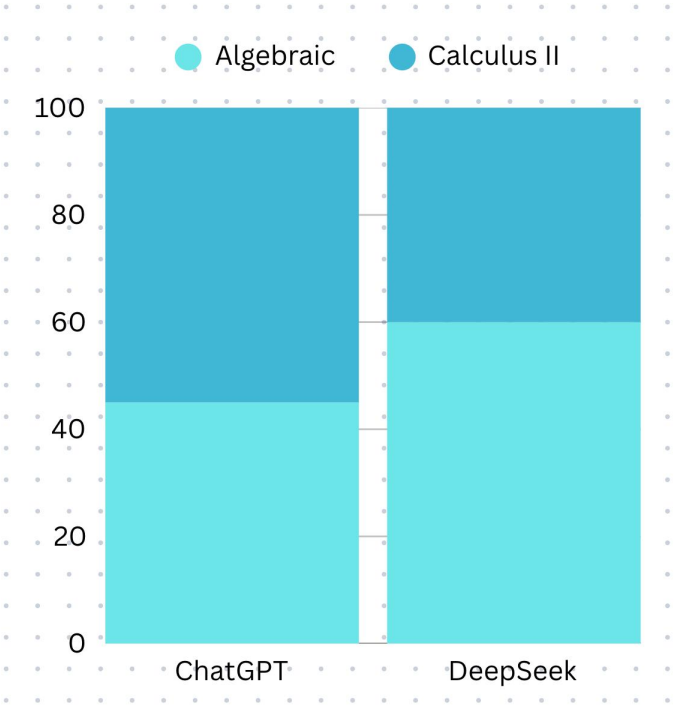


Figure 1: Stacked Bar Chart showing performance of the AI models in Algebraic and Calculus II Problems

Figure 1 presents a stacked bar chart comparing the performance of ChatGPT and DeepSeek on two types of mathematical problems: algebraic and Calculus II. Each bar is divided into two segments, representing the success proportion in each category. ChatGPT performs stronger in Calculus II, while DeepSeek performs slightly better in algebraic reasoning. Ultimately, both models achieved a combined success rate of 100% across all tasks.

2. Ability to Explain Step-by-Step Reasoning
    - ChatGPT was more effective in explaining the relevance of each step, even though it was familiar with complex calculus formulas.
    - DeepSeek was short on thoroughness in expressing why certain steps were taken, especially during transitions between major stages of the solution, although it provided accurate answers.
3. Response Speed

- Both AI chatbots computed the results under 10 seconds. However, ChatGPT's answer appeared slightly faster on average due to its brief solving style. "Users reported high confidence in ChatGPT's ability to provide accurate and immediate responses, attributing its utility to consistent interaction quality and platform reliability." (Mun, I. B. (2025), Online Information Review)

B. Consistency and Reliability

1. Accessibility Across Platforms and Regions

- ChatGPT was accessible via OpenAI and is known for high availability across browsers and platforms.
- DeepSeek was accessed via web-based interfaces and maintained functionality throughout testing, though it occasionally showed interface glitches depending on the browser type.

2. Repeatability in Providing Consistent Solutions

- Both chatbots gave the same correct solutions when the problems were resubmitted, indicating a high consistency level.
- ChatGPT had more stability in structuring its answers, while DeepSeek's layout of the solutions differed across various attempts.

3. Accuracy of Results

- Both AI models correctly computed mathematical results across both problems. Therefore, accuracy was 100% in this sample.

C. Cost of Maintenance

1. Free vs. Paid Versions

- ChatGPT's free versions are enough for basic problem-solving. However, the more advanced models, which are more likely to produce more consistent outputs, are accessed through a paid subscription.
- DeepSeek offered free access to its advanced models, providing it with an advantage for budget-conscious users.

2. Cost with Respect to Efficiency Analysis
    - ChatGPT showed a higher efficiency level in solving problems because of its brief answers. For users familiar with advanced mathematics, this makes a good choice.
    - DeepSeek demonstrated a more educational response with extensive verifications, which makes it suitable for learning contexts, though it is occasionally unclear in some of its step-by-step transitions.

Category	**ChatGPT**	**DeepSeek**
Correctness	✅ Accurate in both problems	✅ Accurate in both problems
Step-by-step explanation	✅ Clear and goal-oriented	⚠️ Detailed but occasionally vague
Method variety	✅ Formula-based (efficient)	✅ Geometric-based (educational)
Response speed	✅ Slightly faster	⚠️ Minor lags noted
Consistency of results	✅ Highly consistent	✅ Mostly consistent
Accessibility	✅ Multi-platform, stable	⚠️ Minor regional/infrastructure gaps
Cost	⚠️ Requires paid access to GPT-4	✅ Free access at time of testing

Table 1: Table Comparing ChatGPT and DeepSeek Across Key Mathematical Metrics

Table 1 summarizes the key findings of our research, comparing ChatGPT and DeepSeek across multiple criteria. While both models were accurate, ChatGPT stood out for its speed, consistency, and efficient methods. On the other hand, DeepSeek offered a more educational

approach, especially in algebraic problems, although it showed minor lags. Accessibility and cost also differed; ChatGPT requires a subscription for its full features, while DeepSeek was freely available during testing.

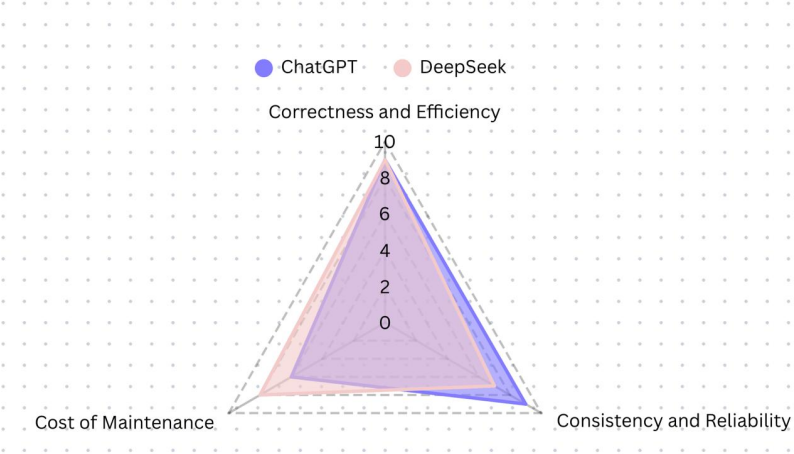


Figure 3: Radar Chart Comparing the Models based on our Key Factors

After discussing and analysing the literary sources, we devised the following radar chart to depict each LLM's correlation with all of our criteria. It can be seen based on Figure 3 that although ChatGPT and DeepSeek are extremely close in correctness and efficiency, ChatGPT outweighs DeepSeek in terms of reliability and consistency of correct answers, while DeepSeek demonstrates that it is cheaper to maintain. It can be suggested that DeepSeek's cheaper nature does not seem to impact its performance much, as it holds its ground as an alternative to ChatGPT.

IV. Conclusion

A. ChatGPT Performance Summary

1. Strengths

ChatGPT showed strong performance across mathematical problems, which demonstrates high efficiency, clarity, and accuracy. Its strength is providing brief and well-structured answers that can be easily understood by users with a good base in mathematical knowledge. This AI model explained its solutions clearly and justified its use of various formulas. Also, its response speed, consistency, and accessibility across platforms through paid versions add to its reliability for practical use.

2. Limitations

Despite its strengths, ChatGPT's concise solutions came at the expense of deeper elaboration. In the algebraic reasoning problem, the chatbot didn't include a detailed verification, which is not ideal for educational purposes. Furthermore, access to the more advanced models requires a paid subscription, which is a barrier to widespread usage.

B. DeepSeek Performance Summary

1. Strengths

DeepSeek demonstrated multiple strengths in methodological diversity, particularly in algebraic reasoning. It solved each problem with a detailed breakdown of solutions, verifying its answers from different angles. In the Calculus II problem, this model embraced a method close to classroom instruction by using vector calculus, which makes it more suitable for students learning new foundational knowledge. The free access to the chatbot also makes it an appealing option for users who want no-cost answers.

2. Limitations

Although DeepSeek's answers were accurate, they sometimes lacked clear transition phases from one major step to another, especially during the high-level stages of the surface area problem. Also, interface instability slightly hindered the user experience.

V. Recommendations

A. Ideal Use Cases for Each AI Model

- ChatGPT is best suited for:
  - Advanced users and professionals looking for brief and accurate results without a step-by-step solution.
  - Time-sensitive contexts: exam preparations and engineering calculations, where brief and short solutions are preferred.
- DeepSeek is best suited for:
  - Students who benefit from a detailed solution and step-by-step explanations.
  - Tutorial-based environment where logic verification is more important than response speed.

B. Potential Improvements for AI-Driven Mathematical Problem-Solving

1. Adaptive Explanation Depth: Using a feature that adjusts explanation depth based on the user's preference and educational level (e.g., beginner, intermediate, professional)
2. Interactive Step-by-Step Method: An interactive problem-solving session, allowing users to ask questions and pause at each step, making these chatbots more educational and beneficial.

3. Unified Visual and Textual Output: Adding dynamic graphs along the explanation can significantly enhance the user's understanding, particularly for algebraic and calculus problems.

VI. References

- Baran, S., & Tong, W. (2024). *50 Shades of AI in Regulatory Science*. VeriSIM Life. https://go.exlibris.link/tSdmh68C

- Blouin, L. (2023). AI's mysterious 'black box' problem, explained. *University of Michigan-Dearborn.* https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained

- Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023). *Mathematical capabilities of ChatGPT*. Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), Datasets and Benchmarks Track. Retrieved from https://ghosts.friederrr.org

- Gibney, E. (2025). China's cheap, open AI model DeepSeek thrills scientists. *Nature (London), 638*(8049), 13-14. https://doi.org/10.1038/d41586-025-00229-6

- Hattenhauer, R. (2024). ChatGPT & Co.: A Workbook for Writing, Research, Creating Images, Programming, and More (1st ed.). Chapman and Hall/CRC. https://doi-org.ezproxy.aub.edu.lb/10.1201/9781003503675

- Katta, K. (2025). Analyzing user perceptions of large language models (LLMs) on reddit: Sentiment and topic modeling of ChatGPT and DeepSeek discussions. https://go.exlibris.link/dWQcCwFX

- Lee, H. (2024). The rise of ChatGPT: Exploring its potential in medical education. *Anatomical Sciences Education, 17*(5), 926-931. https://doi.org/10.1002/ase.2270

- Mun, I. B. (2025). A study of the impact of ChatGPT self-efficacy on the information seeking behaviors in ChatGPT: The mediating roles of ChatGPT characteristics and utility. *Online Information Review, 49*(2), 373-394. https://doi.org/10.1108/OIR-01-2024-0051

- Paul, A. (2023, July 19). ChatGPT's accuracy has gotten worse, study shows. *Popular Science*.https://www.popsci.com/technology/chatgpt-human-inaccurate/

- Nathan Ayele. (2024). The Usage of New AI Technologies by College Students and its Influence on Learning and Dependence [Documents]. Honors Projects: Open Access (2019-present). Caldwell University Archives. https://jstor.org/stable/community.37312334

- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., & Guo, D. (2024). *DeepSeekMath: Pushing the limits of mathematical reasoning in open language models* (arXiv:2402.03300v3). arXiv. https://arxiv.org/abs/2402.03300

VII. Appendix

1. ChatGPT response on algebra

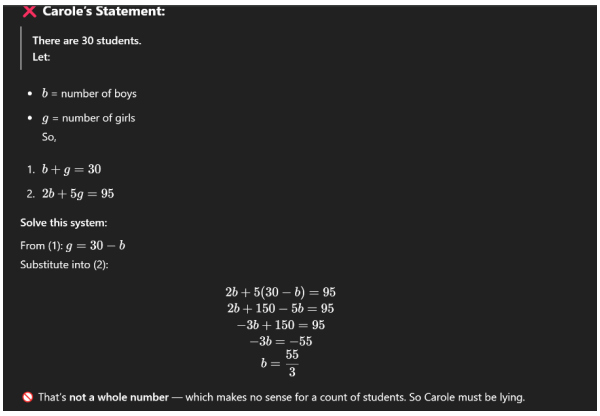


Figure 3: ChatGPT response on algebraic problems

2. DeepSeek response on algebra

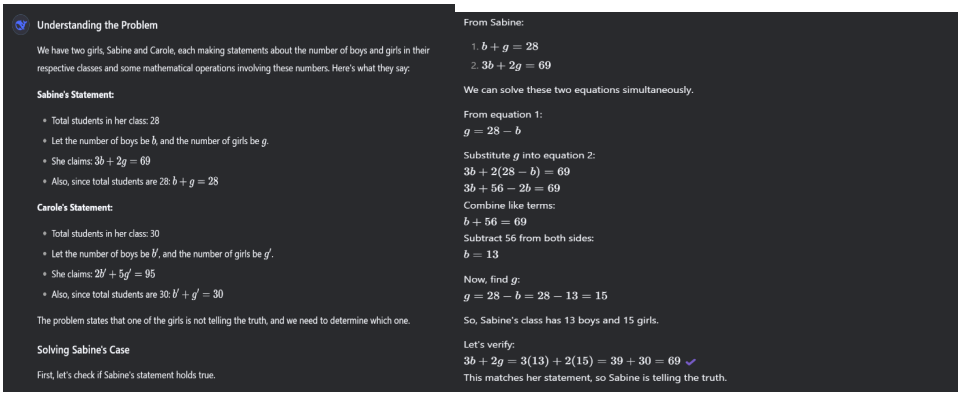


Figure 4: Deepseek response on algebraic problems

3. ChatGPT response on calculus

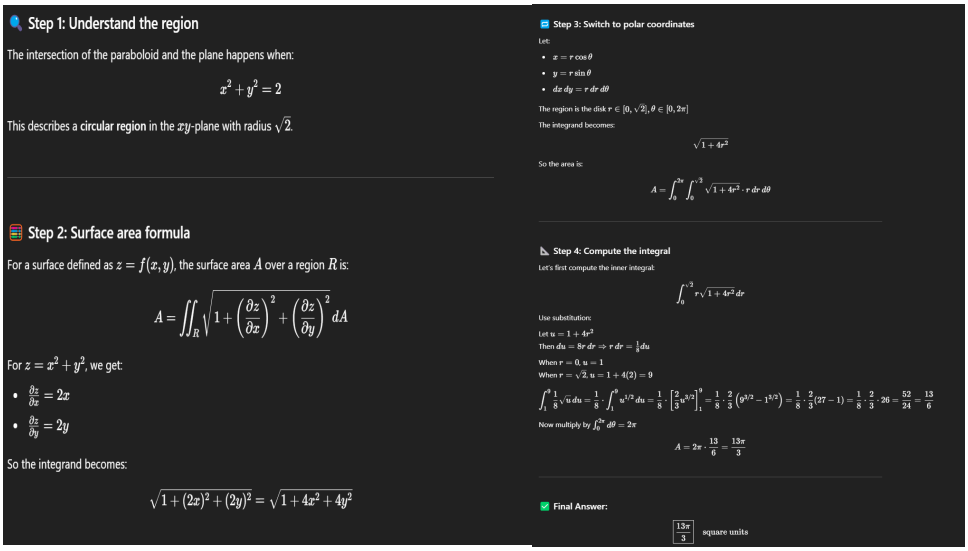


Figure 5: ChatGPT response on surface area problems

4. Deepseek response on calculus

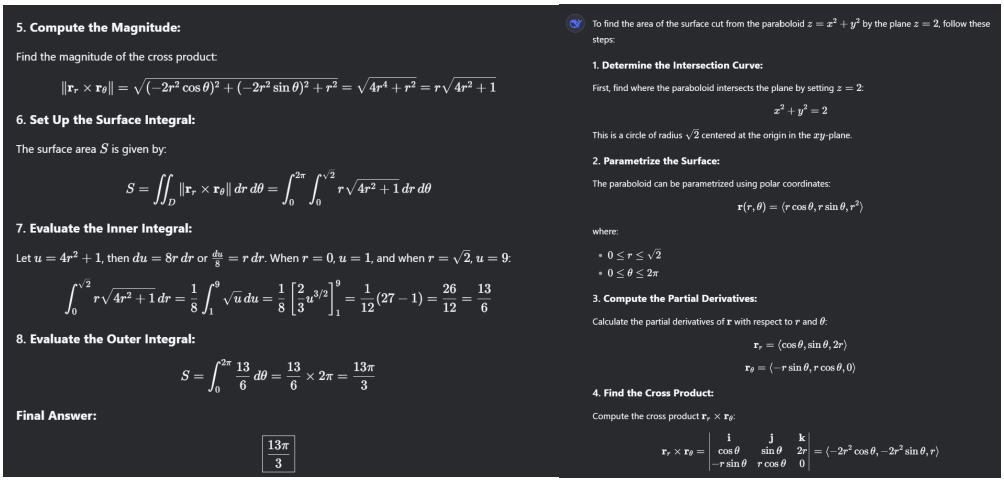


Figure 6: Deepseek response on surface area problems