JOURNAL OF
EMERGING INVESTIGATORS

Article

Comparison of three large language models as middle school math tutoring assistants

Hamsini Ramanathan[1], Ramanathan Palaniappan[1]
[1]Open Window School, Bellevue, WA

SUMMARY

Middle school math forms the basis for advanced mathematical courses leading up to the university level. Large language models (LLMs) have the potential to power next-generation educational technologies, acting as digital tutors to students. The main objective of this study was to determine whether LLMs like ChatGPT, Bard, and Llama 2 can serve as reliable middle school math tutoring assistants on three tutoring tasks: hint generation, comprehensive solution, and exercise creation. Our first hypothesis was that ChatGPT would perform better in completing all three tutoring tasks than Bard and Llama 2 due to its largest model size (175 billion parameters). Our second hypothesis was that Bard would perform better than Llama 2 in generating comprehensive correct solutions due to its relatively higher model size (137 billion parameters) than Llama 2 (70 billion parameters). We curated medium-difficulty, word-based middle school math problems on algebra, number theory, and counting/probability from The Art of Problem Solving and Khan Academy. A human tutor evaluated the LLMs' performance on each tutoring task. Contrary to our first hypothesis, results showed that ChatGPT didn't perform uniformly better than Bard and Llama 2 on all the tasks. ChatGPT outperformed both Bard and Llama 2 only in the comprehensive solution task. Bard didn't perform better than Llama 2 in the comprehensive solution task which does not support our second hypothesis. We conclude that middle school math teachers can use a combination of ChatGPT, Bard, and Llama 2 as assistants based on the specific tutoring task.

INTRODUCTION

An exam administered by the National Assessment of Educational Progress in 2022 showed that a meager 26% of eighth graders were proficient in math, down from 34 percent in 2019 with declining scores in almost every US state (1). Furthermore, math scores began declining in 2012, and average scores are now lower than they were before the pandemic (2). Teacher absenteeism and burnout have been cited as major reasons for this important issue (2-3).

Artificial Intelligence (AI) refers to the development of computer systems that can perform tasks that typically require human intelligence (4). These tasks include learning, reasoning, problem-solving, perception, natural language understanding, and speech recognition (4). Recent advances in large language models like ChatGPT, Bard, and Llama 2 may be able to address the tutoring challenges (5-7). Large language models (LLMs) are a specialized class of AI model that uses natural language processing to understand and generate humanlike text-based content in response (8). Models like ChatGPT are based on a deep learning neural network called a generative pre-trained transformer, which uses large amounts of digital content (large volumes of text from books, Wikipedia articles, social media posts, and other publicly available information) to read and produce humanlike text, making them a valuable tool in automating tasks (9-10). LLMs are trained at predicting the next word using probabilities (11-12). Given a piece of text as input, the LLM generates a sorted list of possible meaningful next words with probability scores for each word and then, picks the word with the highest probability. The LLM repeats this process continually to generate a paragraph of text. LLMs have the potential to enhance math education in several ways, such as by acting as a digital tutor for a student, as a personal assistant for an educator and teacher, or as a digital peer for collaborative learning (13-15).

Recent work has focused on the ability of deep learning models and LLMs to solve complex word-based mathematical problems (16-18). However, these studies purely focus on the models' abilities to solve a math problem correctly. Hence, there is a lack of systematic study that benchmarks the LLMs' performance as a math tutor on a comprehensive set of math education tasks.

The main objective of our study was to determine whether LLMs, like ChatGPT, Bard, and Llama 2, can serve as reliable and effective math tutoring assistants for middle school algebra, number theory, and counting/probability subjects. We excluded geometry from our study since the LLMs are text based and geometry problems require visual input. We chose ChatGPT, Bard, and Llama 2 as these models are freely accessible through their web platforms and well supported and maintained by OpenAI, Google, and Meta, respectively. We curated medium-difficulty, word-based, middle school math problems from sources such as Khan Academy and The Art of Problem Solving.

Past research has studied the effectiveness of LLMs as a programming tutor (19). Based on that work, we considered three common math education tasks normally performed by a human tutor: hint generation – prompting the LLM to generate a hint for a math problem without revealing too much to the student, comprehensive solution – prompting the LLM for a comprehensive step-by-step solution to the problem, and exercise creation – prompting the LLM to generate three similar simpler exercise problems based on an example problem.

Our first hypothesis was that ChatGPT would perform

https://doi.org/10.59720/23-253

No.	Subject	Problem Statement	Source
1	Algebra	Alain throws a stone off a bridge into a river below. The stone's height (in meters above the water), x seconds after Alain threw it, is modeled by: $h(x) = -5x + 10x + 15$ How many seconds after being thrown will the stone hit the water?	Khan Academy
2	Algebra	Find the volume of a tank whose base has an area of $(3x^2 + 30x + 5)$ square feet, and whose height is $(8x - 5)$ feet.	Khan Academy
3	Algebra	A plane has 360 total seats, which are divided into economy class and business class. For every 13 seats in economy class, there are 5 seats in business class. How many seats are in every class?	Khan Academy
4	Algebra	What is the ratio of x to y if $\frac{10x-3y}{13x-2y} = 3/5$?	Art of Problem Solving
5	Algebra	Let $f(x) = 2x - 6$, and $g(x) = 3x - 9$. Find $f(g(2))$.	Art of Problem Solving
6	Counting/ Probability	Michael is taking a quiz in his music history class. The teacher writes the names of 6 songs on the board, and then plays 4 songs out of the 6 one after the other in a random sequence. Michael then needs to identify each of the songs in the sequence they were played. Suppose that Michael hasn't studied and is randomly guessing the songs. What is the probability that Michael correctly identifies all 4 songs in the correct order?	Khan Academy
7	Counting/ Probability	If you flip a coin and roll a 6-sided die, what is the probability that you will flip a heads and roll a 2?	Khan Academy
8	Counting/ Probability	A bag contains hundreds of red, orange, green and blue marbles. Marbles are drawn randomly from the bag, one at a time, and not replaced. How many marbles must be taken out of the bag in order to ensure that at least three are of the same color?	Khan Academy
9	Counting/ Probability	The sundae bar at Sarah's favorite restaurant has 5 toppings: hot fudge, sprinkles, walnuts, cherries, and whipped cream. In how many different ways can Sarah top her sundae if she is restricted to at most 2 toppings?	Khan Academy
10	Counting/ Probability	When flipping a fair coin 7 times, what is the probability that at least 4 heads appear?	Art of Problem Solving
11	Number Theory	Find the smallest positive integer n such that $617n \equiv 943n \pmod{18}$.	Art of Problem Solving
12	Number Theory	What are the possible units digit of a perfect fourth power written in base 5?	Art of Problem Solving
13	Number Theory	Determine the largest possible integer n such that 942! is divisible by 15^n.	Art of Problem Solving
14	Number Theory	What is the smallest positive four-digit number that gives a quotient 432 with remainder 2 when divided by a positive one-digit number?	Art of Problem Solving
15	Number Theory	Find each two-digit number that is equal to four times the sum of its digits.	Art of Problem Solving

Table 1: The curated middle school math problems used in our experiments. 15 total problems covering fundamental concepts were selected from the primary middle school math subjects: algebra, counting/probability, and number theory. Geometry was excluded as the LLMs were restricted to text input.

better than Bard and Llama 2 on all the tutoring tasks due to its large model size of 175 billion parameters which was the best available model (20). Bard is based on the LaMDA family of models, which have up to 137 billion parameters (21). For Llama 2, we used the 70 billion parameter model which was the largest available model. Our second hypothesis was that Bard would perform better than Llama 2 in generating comprehensive correct solutions due to its relatively higher model size (137 billion parameters) compared to Llama 2 (70 billion parameters). ChatGPT did not outperform Bard and Llama 2 across all three tutoring tasks which is contrary to our first hypothesis. ChatGPT and Llama 2 outperformed Bard on the hint generation task. ChatGPT was the clear winner in the comprehensive solution task. There was no clear winner in the exercise creation task. Contrary to our second hypothesis, both Bard and Llama 2 performed poorly in generating comprehensive correct solutions with no model being better than the other.

You are a math tutor. I am trying to solve the below problem and I am stuck. Can you provide me with a minimal hint without solving the full problem?

{{ problem statement }}

Figure 1. Hint generation prompt. Prompt sent to the LLM to generate hints for the math problem stated within {{problem statement}}.

RESULTS

To conduct the experiments, we chose five problems for each math topic covered in middle school: algebra, number theory, and counting/probability – 15 problems in total. We sourced the problems from The Art of Problem Solving courses and Khan Academy videos (22-23). We selected 15 problems to ensure good representations of the core middle school math concepts (**Table 1**). A human tutor with access to the solutions scored the tasks and were provided with the model responses but the tutor was not aware of the model that they were scoring. We converted the scores (range of 0 to 1) into percentages.

Hint generation

The input to the LLMs to test their hint generation ability consisted of a math problem and a detailed prompt asking the model for an assistive hint to help solve the problem accurately (**Figure 1**). This task had three quality attributes: correctness, comprehensibility, and concealment.

The "correctness" attribute captured if the hint could help the student proceed in the right direction. The "comprehensibility" attribute captured if the hint was easily understandable by an average student. The "concealment" attribute measured if the model's hint did not reveal too much about the solution.

The overall average scores were as follows: 100% for ChatGPT, 58% for Bard, and 84% for Llama 2 (**Figure 2a**). ChatGPT scored 100% in all the subject areas. Bard scored 53% in algebra, 47% in counting/probability, and 73% in number theory. Llama 2 scored 100% in algebra, 93% in counting/probability, and 60% in number theory outperforming Bard in algebra and counting/probability convincingly (**Figure 2a**).

ChatGPT scored 100% across all quality attributes (**Figure 2b**). Bard scored 73% in correctness, 80% in comprehensibility, and 20% in concealment. Llama 2 scored 83% in correctness, 77% in comprehensibility, and 93% in concealment displaying a consistent performance across all quality attributes (**Figure 2b**).

Across all the math problems, ChatGPT and Llama 2 showed significantly better overall average scores than Bard

https://doi.org/10.59720/23-253

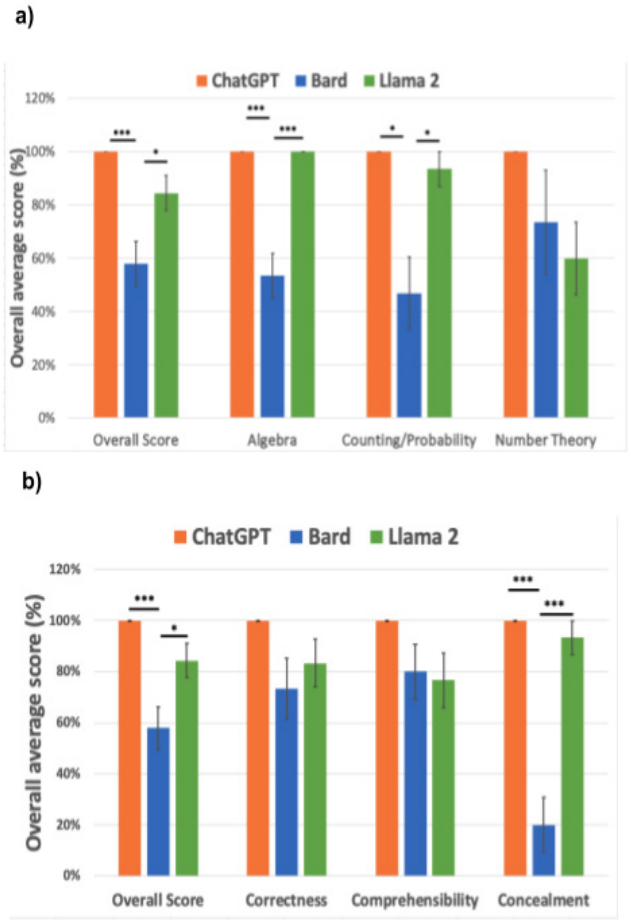

Figure 2. Performance of the LLMs on the hint generation task. a) Bar graph comparing the overall score on all problems and comparing the LLMs' performance by each subject area (***p < 0.001, *p < 0.05). b) Bar graph comparing the overall score on all problems and comparing the LLMs' performance by each task quality attribute (***p < 0.001, *p < 0.05). Error bars represent standard error.

on the hint generation task ($p < 0.05$, one-way ANOVA with a post-hoc Tukey-HSD; **Figure 2a**). Both ChatGPT and Llama 2 performed significantly better than Bard on algebra and counting/probability ($p < 0.05$, one-way ANOVA with a post-hoc Tukey-HSD; **Figure 2a**). Similarly, ChatGPT and Llama 2 performed significantly better than Bard on the concealment quality attribute ($p < 0.001$, one-way ANOVA with a post-hoc Tukey-HSD; **Figure 2b**). In our experiments, Bard revealed the solutions to most of the problems when asked for a hint thus lowering Bard's score on concealment.

Comprehensive solution

The input to the LLMs to test their ability to come up with a comprehensive solution consisted of a math problem and a detailed prompt asking the model for a clear step-by-step solution to the math problem (**Figure 3**). This task had two quality attributes: correctness and comprehensibility. The "correctness" attribute captured if the LLM solved the problem correctly and the "comprehensibility" attribute captured if the solution logically contained all the necessary steps leading to the final correct answer.

The overall average scores were as follows: 82% for ChatGPT, 37% for Bard, and 13% for Llama 2 (**Figure 4a**). ChatGPT scored 80% in algebra, 95% in counting/probability, and 70% in number theory displaying a consistent performance. Bard scored 40% in algebra, 40% in counting/probability, and 30% in number theory. Llama 2 scored 20% in algebra, 20% in counting/probability, and 0% in number theory displaying the worst performance of all the three models (**Figure 4a**).

You are a math tutor. I am stuck on a problem, can you please provide a step by step explanation for the problem below to help me out.

{{ problem statement}}

Figure 3. Comprehensive solution prompt. Prompt sent to the LLMs to generate a comprehensive step-by-step solution for the math problem stated within {{problem statement}}.

On the quality attributes, ChatGPT scored 87% in correctness and 77% in comprehensibility (**Figure 4b**). Bard scored 40% in correctness and 33% in comprehensibility.

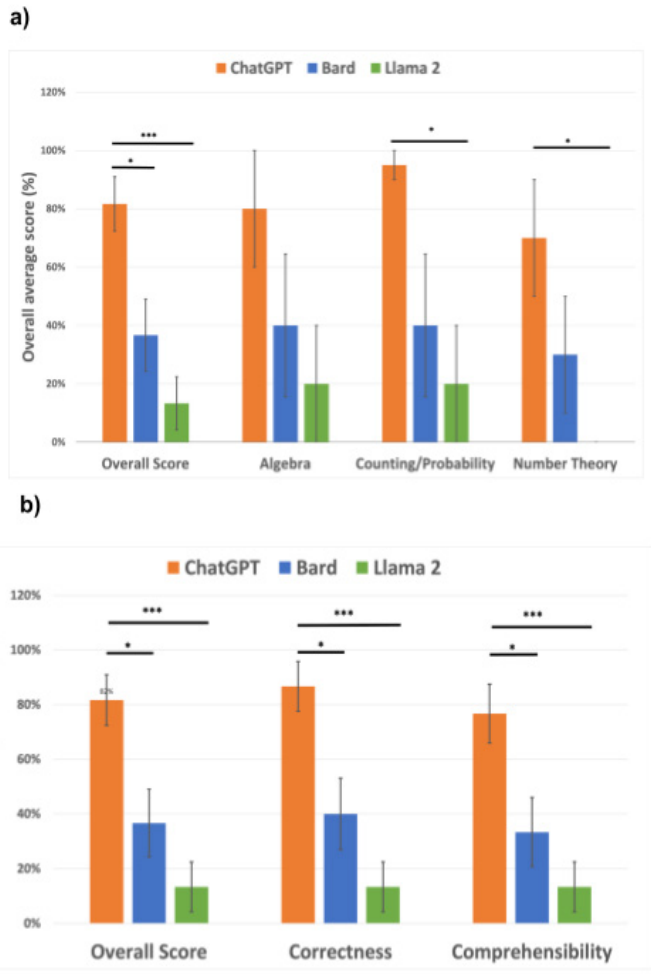

Figure 4. Performance of the LLMs on the comprehensive solution task. a) Bar graph comparing the overall score on all problems and comparing the LLMs' performance by each subject area (***p < 0.001, *p < 0.05). b) Bar graph comparing the overall score on all problems and comparing the LLMs' performance by each task quality attribute (***p < 0.001, *p < 0.05). Error bars represent standard error.

https://doi.org/10.59720/23-253

You are a math tutor. I am stuck on the problem below, hence could you please provide me with 3 exercise problems that utilize the same concepts as the ones included in the problem below, but of simpler difficulty for practice.

{{ problem statement }}

Figure 5. Exercise creation prompt. Prompt sent to the LLMs to generate three similar and simpler exercise problems for the math problem stated within {{problem statement}}.

Llama 2 scored 13% in correctness and 13% in comprehensibility (**Figure 4b**).

Across all the math problems, ChatGPT showed significantly better overall average scores than Bard and Llama 2 on the comprehensive solution task ($p < 0.05$, one-way ANOVA with a post-hoc Tukey-HSD; **Figure 4a**). ChatGPT performed significantly better than Llama 2 on counting/probability and number theory ($p < 0.05$, one-way ANOVA with a post-hoc Tukey-HSD; **Figure 4a**). ChatGPT performed significantly better than Bard on both quality attributes ($p < 0.05$, one-way ANOVA with a post-hoc Tukey-HSD; **Figure 4b**). Similarly, ChatGPT performed significantly better than Llama 2 on both quality attributes ($p < 0.001$, one-way ANOVA with a post-hoc Tukey-HSD; **Figure 4b**).

Exercise creation

The input to the LLMs to test their exercise creation ability consisted of a math problem and a detailed prompt asking the model to create three similar simpler versions of the problems for the student to practice (**Figure 5**). This task had two quality attributes: correctness and simpler. The "correctness" attribute captured if the LLM generated three similar exercise problems. The "simpler" attribute captured if the three problems were of similar or simpler complexity.

The overall average scores were as follows: 70% for ChatGPT, 83% for Bard, and 73% for Llama 2 (**Figure 6a**). ChatGPT scored 80% in algebra, 50% in counting/probability, and 80% in number theory. Bard scored 90% in algebra, 80% in counting/probability, and 80% in number theory. Llama 2 scored 70% in algebra, 70% in counting/probability, and 80% in number theory (**Figure 6a**).

On the quality attributes, ChatGPT scored 87% in correctness and 53% in simpler (**Figure 6b**). Bard scored 93% in correctness and 73% in simpler. Llama 2 scored 100% in correctness and 47% in simpler (**Figure 6b**). ChatGPT's low overall score may be largely attributable to its low score in the simpler attribute, which was 53%.

Across all the math problems, none of the three models showed significantly better overall average scores relative to each other on the exercise creation task ($p < 0.05$, one-way ANOVA with a post-hoc Tukey-HSD; **Figure 6a**).

DISCUSSION

Middle school students' proficiency in fundamental math concepts is paramount as it lays the groundwork for their mathematical journey. Leveraging the capabilities of LLMs can significantly enhance the effectiveness of math tutoring strategies for these students. In this study, we compared the tutoring abilities of three LLMs — ChatGPT, Bard, and Llama 2 — by testing their ability to solve math problems,

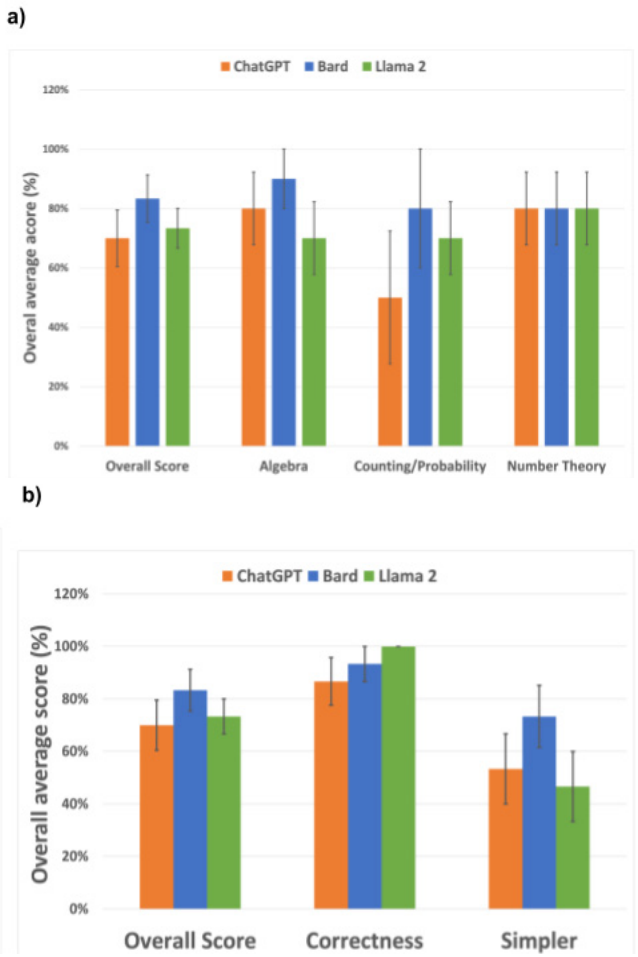

Figure 6. Performance of the LLMs on the exercise creation task. a) Bar graph comparing the overall score on all problems and comparing the LLMs' performance by each subject area. b) Bar graph comparing the overall score on all problems and comparing the LLMs' performance by each task quality attribute. Error bars represent standard error.

provide hints, and create additional practice problems. Our first hypothesis was that ChatGPT would perform better than Bard and Llama 2 on all the three tutoring tasks. Our second hypothesis was that Bard would perform better than Llama 2 in generating comprehensive correct solutions.

The findings compel us to challenge our initial hypotheses. We observed a nuanced interplay with different LLMs exhibiting commendable performance in distinct tasks. Consequently, instead of designating a single LLM as the superior choice, a blended approach involving all three LLMs based on the tutoring task emerges as an optimal strategy. ChatGPT and Llama 2 prove to be a better bet in hint generation (**Figure 2a**). In our experiments, Bard revealed the solutions to most of the problems when asked for a hint. ChatGPT is the clear winner in the comprehensive solution task (**Figure 4a**). Bard and Llama 2 struggled to generate correct solutions. In exercise creation, all three LLMs perform well with no clear winner based on statistical tests (**Figure 6a**). We conclude that ChatGPT, Bard, and Llama 2 can serve as middle school math tutoring assistants to teachers based on specific tutoring tasks where each LLM excels. LLMs can

https://doi.org/10.59720/23-253

be unpredictable and are constantly evolving; hence, we recommend that teachers and educators leverage them as personal digital assistants in their classrooms (24).

This study emphasized the pivotal role of LLMs in augmenting middle school math tutoring, demonstrating good levels of efficacy. It is essential to remember that our study encompassed only 15 math problems, with 5 from each subject area (algebra, probability, and number theory). For broader applicability in everyday tutoring scenarios, rigorous validation of this study is necessary. Given our limited dataset and the constraints of only 15 problems, we may have yet to cover all the familiar concepts, ideas, and methodologies within the chosen subjects. Future studies can focus on building a more extensive problem dataset comprising hundreds of problems. The new dataset should include diverse problem types, variations in problem-wording, and numerous problems covering various concepts within a given sub-topic like quadratic equations, greatest common divisor, least common multiple, functions, and inequalities. Additionally, our current study used a human tutor to grade the responses. Future studies could also leverage multiple evaluators, including middle school teachers and students, to grade the LLM responses, which would be more representative of real-world scenarios.

While the current study focused on readily available LLMs from the tech companies, future studies can focus on evaluating a larger curated problem dataset across other open-source LLMs. We can also develop fine-tuning techniques to improve the performance of current LLMs like Bard and Llama 2 as better middle school math tutors (25-28). LLMs can also help improve patient outcomes by leveraging the availability of vast patient data and medical literature in the healthcare industry (29).

LLMs can understand human language, solve math problems, and act as a digital tutoring assistant. They can help automate, or semi-automate, many day to day tasks that teachers do, thereby freeing up teachers' time for personalized instruction to students. Our results indicate that a combination of ChatGPT, Bard, and Llama 2 can generate meaningful responses to the most common tutoring tasks on hint generation, comprehensive solution, and exercise creation, which can ease the load on teachers. Given this new technological milestone in AI, we hope our study will advance further research and assist teachers in helping students pursue mathematical excellence.

MATERIALS AND METHODS

Problems Dataset

To conduct the study, we picked three core subject areas that are commonly taught in middle school: algebra, number theory and counting/probability. Next, we curated five problems for each math topic. We selected problems that best covered the fundamental concepts in each specific math topic. We brainstormed the main tutoring tasks of math tutors: a) providing a hint to the student, b) showing a clear step by step solution, and c) providing new exercise problems to practice. Based on this breakdown, we conducted experiments to evaluate each of the tutoring tasks for each LLM.

LLM Versions

For the experiments, we accessed the latest version of the LLMs via their freely available web platforms between the weeks of Sep 10th, 2023 and Sep 20th, 2023 (25-27). ChatGPT's underlying model in the OpenAI chat interface was GPT-3.5. We used the latest version of Bard through its web platform. For Llama 2, we used the 70 billion parameter model through its web platform with a default recommended temperature setting of 0.75. Temperature refers to the randomness of the model output.

Experiment Procedure

For each LLM, we evaluated three tutoring tasks to accommodate different instruction techniques that would be performed by a human math tutor: a) Hint generation– prompting the LLM to generate a hint for a math problem without giving away too much to the student; b) comprehensive solution– prompting the LLM for a comprehensive step-by-step solution to the problem; c) exercise creation–- prompting the LLM to generate three similar straightforward exercise problems based on an example problem.

We modified the prompt template for each tutoring task by replacing the placeholder {{problem statement}} with one of the 15 math problems (**Table 1**). The LLM was then invoked with the modified prompt through its web platform. The exact result from the LLM was recorded and evaluated by a human tutor who had access to the solutions. We repeated this process for every combination of math problem (15 problems), LLM (3 models), and tutoring task (3 tasks), resulting in 135 different trials that were independently evaluated and scored by the human tutor.

Finally, the LLM generated response was assigned values of 0 or 1 by a human tutor along the task specific quality attributes. Only binary values were used, and 1 was better than 0. The human tutor was well versed in middle school math and had access to all the math problem solutions.

Hint generation task evaluation

The prompt sent to the LLM was: "You are a math tutor. I am trying to solve the below problem and I am stuck. Can you provide me with a minimal hint without solving the full problem? {{problem statement}}." This task has three quality attributes: correctness, comprehensibility, and concealment. The "correctness" attribute captures if the hint can help the student proceed in the right direction. It is set to 0 if the response could lead the student in an incorrect direction. The "comprehensibility" attribute captures if the hint is easily understandable by an average student and is set to 0 if it contains unclear or unrelated information. The "concealment" attribute measures if the model's hint does not reveal too much about the solution – it is set to 0 if the model's hint response reveals more information than needed about the solution.

Comprehensive solution task evaluation

The prompt sent to the LLM was: "You are a math tutor. I am stuck on a problem, can you please provide a step by step explanation for the problem below to help me out. {{problem statement}}." This task has two quality attributes: correctness and comprehensibility. The "correctness" attribute captures if the LLM solved the problem correctly – 1 if the solution is correct, otherwise 0. The "comprehensibility" attribute captures if the solution contains all the necessary steps leading to the final answer. We assign 0 if the solution has any logical flaws (even though the final answer may be correct).

https://doi.org/10.59720/23-253

Exercise creation task evaluation

The prompt sent to the LLM was: "You are a math tutor. I am stuck on the problem below, hence could you please provide me with 3 exercise problems that utilize the same concepts as the ones included in the problem below, but of simpler difficulty for practice. {{problem statement}}." This task has two quality attributes: correctness and simpler. The "correctness" attribute captures if the LLM generated three similar problems – we assign 0 if any of the problems deviated from the original problem type provided in the prompt. The "simpler" attribute captures if the three problems are of similar or simpler complexity. We assign 0 if any of the problems is more complex than the original problem.

Statistical Analyses

One-way ANOVA with post-hoc Tukey-HSD tests were conducted to evaluate statistical differences in overall average scores at a significance level of p=0.05.

Received: October 1, 2023
Accepted: December 5, 2023
Published: May 2, 2024

REFERENCES

1. "Math Scores Fell in Nearly Every State, and Reading Dipped on National Exam." *The New York Times*. www.nytimes.com/2022/10/24/us/math-reading-scores-pandemic.html. Accessed 02 Sept. 2023.
2. "Middle Schoolers' Reading and Math Scores Plummet." *Axios.com*. Accessed 04 Sept. 2023.
3. "Quantifying an alarming teacher shortage." *Axios.com*. www.axios.com/2022/09/13/natiional-teacher-shortage-burnout-pandemic-education-deficit. Accessed 04 Sept. 2023.
4. "What Is Artificial Intelligence (AI)?" *IBM*. www.ibm.com/topics/artificial-intelligence. Accessed 24 Feb. 2024.
5. "ChatGPT." *OpenAI*. openai.com/blog/chatgpt. Accessed 04 Sept. 2023.
6. "Bard." *Google*. bard.google.com. Accessed 04 Sept. 2023.
7. "Llama 2." *Meta AI*. www.ai.meta.com/llama. Accessed 04 Sept. 2023.
8. "Introduction to Large Language Models | Machine Learning | Google for Developers." *Google*. developers.google.com/machine-learning/resources/intro-llms. Accessed 24 Feb. 2024.
9. "Explained: Neural Networks." *MIT News | Massachusetts Institute of Technology*. news.mit.edu/2017/explained-neural-networks-deep-learning-0414. Accessed 04 Sept. 2023.
10. "Improving Language Understanding by Generative Pre-Training." *OpenAI*. cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. Accessed 04 Sept. 2023.
11. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017). https://doi.org/10.48550/arXiv.1706.03762
12. "What Is Chatgpt Doing ... and Why Does It Work?" *Stephen Wolfram Writings RSS*. writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work. Accessed 04 Sept. 2023.
13. 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. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4337484
14. Lim, Weng Marc, et al. "Generative AI and the Future of Education: Ragnarök or Reformation? A Paradoxical Perspective from Management Educators." *The International Journal of Management Education*, vol. 21, no. 2, July 2023, p. 100790, https://doi.org/10.1016/j.ijme.2023.100790
15. "Khanmigo Education AI Guide." *Khan Academy*. www.khanacademy.org/khan-labs. Accessed 04 Sept. 2023.
16. Lu, Pan, et al. "A survey of deep learning for mathematical reasoning." *arXiv* preprint arXiv:2212.10535 (2022), https://doi.org/10.48550/arXiv.2212.10535
17. Peng, Shuai, et al. "MathBERT: A pre-trained model for mathematical formula understanding." *arXiv preprint* arXiv:2105.00377 (2021), https://doi.org/10.48550/arXiv.2105.00377
18. "GPT-4." *OpenAI*. openai.com/research/gpt-4. Accessed 04 Sept. 2023.
19. Phung, Tung, et al. "Generative AI for Programming Education: Benchmarking ChatGPT, GPT-4, and Human Tutors." *International Journal of Management* 21.2 (2023): 100790. https://doi.org/10.48550/arXiv.2306.17156
20. "OpenAI Platform." *OpenAI*. platform.openai.com/docs/model-index-for-researchers. Accessed 21 Nov. 2023.
21. Thoppilan, Romal, et al. "Lamda: Language models for dialog applications." *arXiv preprint* arXiv:2201.08239 (2022), pp 2-4, https://doi.org/10.48550/arXiv.2201.08239
22. "Challenge Your Student to Reach Their Fullest Academic Potential." *Art of Problem Solving*. artofproblemsolving.com. Accessed 04 Sept. 2023.
23. "Free Online Courses, Lessons & Practice." *Khan Academy*. www.khanacademy.org. Accessed 04 Sept. 2023.
24. "The Unpredictable Abilities Emerging from Large AI Models." *Quanta Magazine*. www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316. Accessed 04 Sept. 2023.
25. "ChatGPT." *OpenAI*. chat.openai.com. Accessed 04 Sept. 2023.
26. "Bard." *Google*. bard.google.com. Accessed 04 Sept. 2023.
27. "Chat with Llama 2." *Replicate*. www.llama2.ai. Accessed 04 Sept. 2023.
28. "Falcon LLM." *Technology Innovation Institute*. falconllm.tii.ae. Accessed 04 Sept. 2023.
29. "Large language models in healthcare: Examples & 10 use cases in 2024." *AIMultiple*. research.aimultiple.com/large-language-models-in-healthcare/. Accessed 28 Jan. 2024