Comparative Analysis of LLMs for Math Education

Isaac Gasparri
Department of Physics, Computer Science, and Engineering
Houghton University
Houghton, NY, USA
isaac.gasparri26@houghton.edu

***Abstract*—Artificial Intelligence (AI) has emerged as a powerful tool in education, with Large Language Models (LLMs) offering significant opportunities to support student learning and educator instruction. This study investigates the comparative performance of four LLMs in solving math problems across three key domains: Algebra, Calculus, and Statistics. A data set of 30 questions, ranging in complexity and domain, was used to evaluate the accuracy, clarity, and valuability of each model's responses. Evaluations were conducted by a human evaluators familiar with the subjects.**

Results revealed that while all models performed similarly in terms of accuracy, the other two categories highlighted performance gaps. Meta AI emerged as the top performer, excelling in clarity and value, particularly in Calculus and Statistics. Julius AI demonstrated strength in solving complex word problems, whereas Gemini excelled at simpler problems. ChatGPT, however, consistently underperformed in all metrics, making it less reliable as a math resource. These findings offer insights into the capabilities of AI tools for math education, equipping educators and students with guidance on model selection based on the type and domain of the problem.

***Index Terms*—artificial intelligence, computers, evaluation framework, model**

I. Introduction

Artificial intelligence (AI) is a growing technology that is widely used throughout the world. 77 percent of devices in use feature some sort of AI functionality [1]. The global AI market is valued at more than 279 billion dollars [2]. 30 percent of Americans had high awareness of the use of AI in their lives, while another 38 percent had at least a medium level of awareness [3]. A lot of this awareness comes from the recent introduction of free-to-use AI chat bots such as OpenAI's ChatGPT and Google's Gemini. These Large Language Models (LLMs) provide users with quick responses to nearly any prompt submitted, but not without some faults; the most concerning being hallucinations. Hallucinations are LLM outputs that are either nonsensical or untrue and remain "a critical challenge that impedes the practical application of LLMs" [4]. However, still about 39 percent of people are willing to trust AI [5].

An area that has seen an increase in the use of AI technology is education. 19 percent of teens who are aware of ChatGPT have said that they have used it for schoolwork [6]. Although there are certainly concerns from teachers and educators about over-reliance on AI, struggling students could actually benefit from the ability to receive help from an AI source. Almost 96 percent of students have access to a smartphone, and 91 percent said they have access to great or okay internet quality at home [7]. With the ease of access to LLMs today, almost every student has the ability to connect with an AI resource to help them learn. In the aftermath of the Covid-19 pandemic, scores in both reading and mathematics decreased significantly. Reading scores dropped by 5 points, the largest since 1990, and math scores dropped for the first time ever, by 7 points [8]. The gap between different demographics of students also widened over that time. Schools with higher diversity of students and in areas of higher poverty levels experienced similar drops to the national average yet the gap, especially in math scores, widened even as scores started to rebound between 2021 and 2022 [9]. However, LLMs have been shown to be potentially useful resources to help both students learn better and educators teach more effectively. But as more and more companies start creating public LLMs, which ones are actually worth investing the time into to make them suitable for use in education?

This study compares four different modern LLMs: ChatGPT, Gemini, Meta AI, and Julius AI, a math-specific model that uses a combination of GPT and Claude models, and their ability to produce accurate and comprehensible solutions to math problems from a variety of subjects students might encounter.

II. Related Works

A survey examining the advancements being made in the ability for LLMs to be used in math education found that LLMs help develop critical thinking and are great for conversational learning. It also found some challenges, the biggest one being misinterpretation of queries, and giving inaccurate responses as a result [10]. Different LLMs were tested in their ability to grade student submissions, and scored above 60 percent in both tested datasets, while human graders only scored around 10 percent better [11]. Another study tested three LLMs in their ability to provide hints, generate solutions, and create additional related problems based on an inputted math question, and found ChatGPT to be a very helpful tool, scoring high in all three categories [12]. A third study compared different versions of GPT Models and showed that newer models performed better on math problems, which is a great sign for the future of AI [13]. LLM technology was used to create single equations that were passed to an external solver to get results similar and often better than other methods that were tested [14].

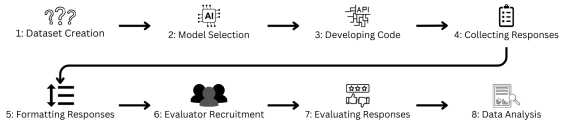

Fig. 1. Methodology Overview

III. Methodology

1 above contains a visual overview of the methodology.

A. Dataset Creation

To start, a new dataset of thirty questions was created using both online open-access textbooks and the ALGEBRA dataset used in Reference [13]. The questions came from three distinct domains in mathematics: Algebra, Calculus, and Statistics. These three in particular were chosen for their importance to students growth. Algebra is one of the first major math classes students take in high school, calculus is very important as both an early college class or a senior year challenge, and statistics has value outside of mathematics as well. 10 questions per domain were selected, with 5 of each being simpler questions, akin to what could be found in a multiple choice exam. The other 5 were more difficult, either containing multiple parts or more explanations, like would be found in a word problem based exam.

B. Model Selection

To select the LLMs to be chosen for testing, some time was spent researching the different available models on the market. Models were chosen based on popularity, ease of use, and specialization in the topic. The four LLMs selected were OpenAI's ChatGPT, Google's Gemini, Meta AI, and Julius AI, an LLM trained specifically for solving math problems. All 4 models have free access to the chatbot, though an account is required to sign in.

C. Developing Code to Generate Answers

Due to the large volume of questions, a code was written to take the list of questions from a spreadsheet, run it through the model using available APIs, and output the responses to another spreadsheet. This method was used for ChatGPT and Gemini. These scripts can be found on GitHub, as well as the dataset used for the project: Comparative-Analysis-of-LLMs-for-Math-Education

D. Collecting Responses

The scripts were run to collect the ChatGPT and Gemini responses, and the Julius and Meta responses were collected using the online chatbots. All of the responses were recorded in a spreadsheet.

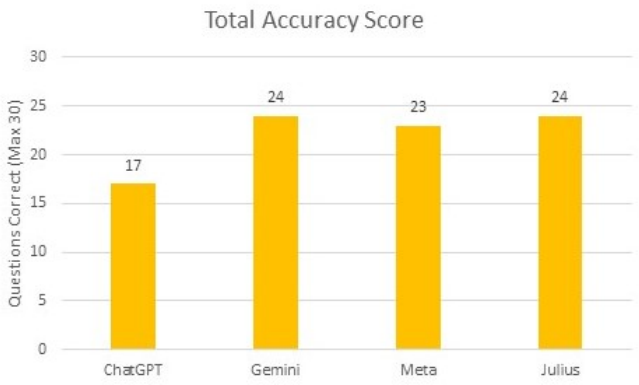

Fig. 2. Overall Accuracy Scores, out of 30 questions

E. Formatting Responses for Evaluation

The questions, correct answers, and responses from each model were inputted into a slideshow presentation to make it easy for the evaluators to compare the responses. As seen on GitHub, the question and answer appeared at the top, and each response was placed in a quadrant below. To avoid any potential bias, the names of the models were not included with the responses, and the quadrant the responses were placed in was randomized.

F. Evaluator Recruitment

5 evaluators were used in the project. 3 were Computer Science majors at Houghton University, 2 juniors and 1 sophomore, who were all in the Machine Learning class in the semester this study was done. The other 2 evaluators were Houghton University seniors in the education department, both of which had concentrations in mathematics. This ensured that all evaluators had familiarity with the topics involved in this research.

G. Evaluation of Responses

1 evaluator was given the task of grading the accuracy of the models, using a simple binary scale: 0 for a wrong answer, 1 for a full correct answer. Then all 5 of the evaluators rated the LLM responses against one another in two different categories: Clarity and Valuability. Descriptions of these and the entire set of instructions given to the evaluators can be found in IX. Thus, each response was given between a 1 and 4 for each question. These were inputted into an anonymous Google Form.

H. Data Analysis

Once all of the grades were collected, each model's response was given an average score for each question, and an overall average score was given by domain and by type, as seen below.

IV. Results

A. Accuracy

None of the models were able to achieve 100 percent accuracy through 30 questions. Julius and Gemini each had 24

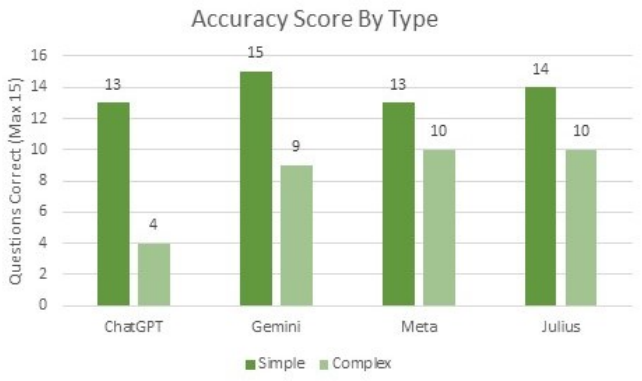

Fig. 3. Accuracy Scores by Question Type, out of 15

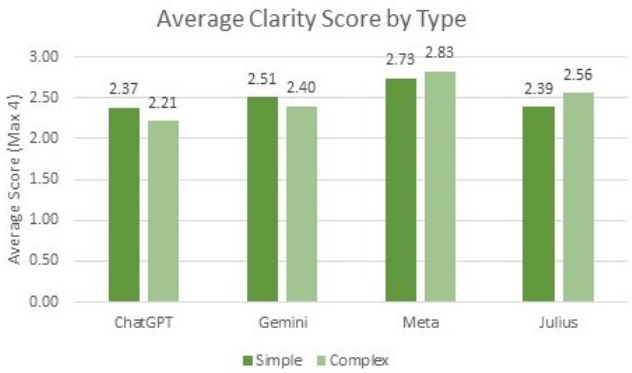

Fig. 5. Average Clarity Scores by Question Type

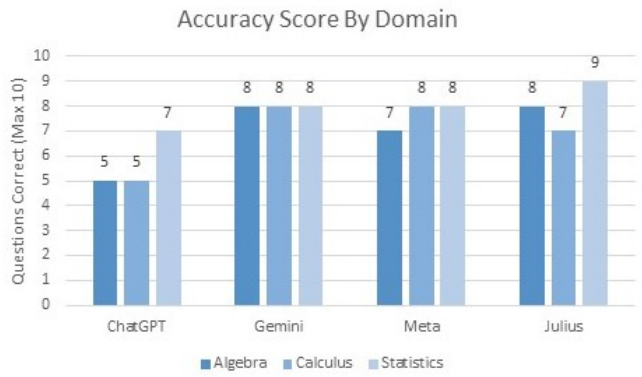

Fig. 4. Accuracy Scores by Question Domain, out of 10

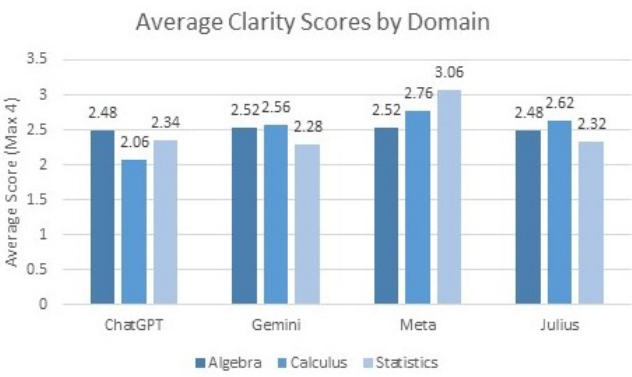

Fig. 6. Average Clarity Scores by Question Domain

correct answers, and Meta was just behind, with 23. ChatGPT struggled the most, only scoring 17 out of 30, or just around 57 percent, a failing grade for most students. See 2 for a visual representation.

Breaking the data down by separate question types, we can see where the models missed most of their questions. See 3. Gemini was actually able to score 15 out of 15 in the simpler questions, but all 4 models decreased in accuracy when asked the more complex questions.

There was not a significant difference between how the models scored per domain, as seen in 4

B. Clarity

The clarity category is where the data begins to get more interesting. Meta ended up scoring the highest of all the models in both simple and complex questions, averaging outstanding average scores of 2.73 and 2.83, respectively. Julius also proved to be clearer when explaining more complex problems compared to the other models, while ChatGPT and Gemini struggled more in that category. See 5 for the whole graph.

Another interesting observation can be found when viewing by domain, where all the models scored within 0.04 points of one another on the Algebra questions (see 6). Meta dominated in statistics, however, scoring an average of 3.06, which was the highest single average score in any category by over 0.2 points.

C. Valuability

Similar to clarity, Meta and Julius were deemed the most valuable responses more often for the problems with higher complexity, but Gemini proved to be just as good as Meta in the simpler category. ChatGPT, as has been across the board, still performed the worst out of the tested models, as seen in 7

By looking at domain, we can see that Gemini was the most valuable model in explaining Algebraic questions, while Meta still dominated in calculus and statistics, as seen in 8

V. Conclusions

In terms of accuracy, Gemini, Julius, and Meta are all on a similar level, regardless of the difficulty of the question or domain. However, Meta clearly shines through the competition in the value and clarity of its explanations of problems. While all of the models are pretty close in algebra, Meta dominated in both calculus and statistics. Thus, we can conclude that Meta should be used in those domains of math. When consulting more difficult word problems, Julius is also a viable option,

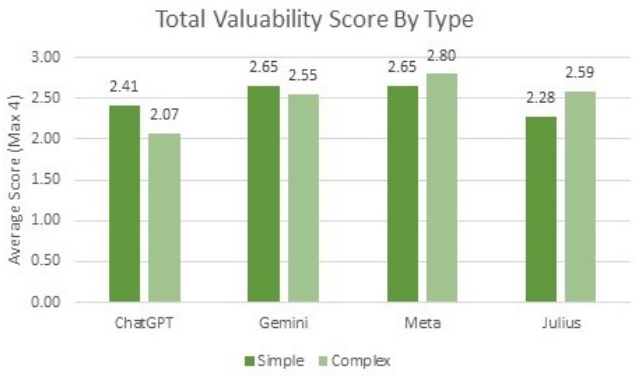

Fig. 7. Average Valuability Scores by Question Type

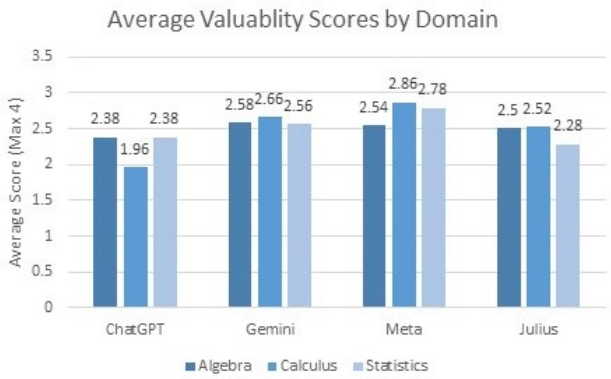

Fig. 8. Average Valuability Scores by Question Domain

while Gemini keeps pace on simpler questions. ChatGPT, however, should be avoided as a math tool, as it almost always scored lower than any of the other models.

VI. Significance of Study

With so many new opportunities being created by AI, it is important to find the best ways to use all of the different features. While not an exhaustive project, this research helps narrow down some options in a little corner of the world. Now, students and educators can be more informed on the capabilities of AI in mathematics as a whole, as well as where to turn to when additional support is needed.

VII. Suggested Improvements for Further Study

Though a much larger time commitment for any human evaluators, having a larger question dataset would definitely improve the study. Additionally, including more domains of mathematics would be interesting as well. Geometry is one in particular that sticks out, as many chatbots now have the ability to input and comprehend images, which would be ideal for solving a lot of geometric problems. Proofs are also a difficult concept for many students, so testing models, especially in their clarity when giving proofs, would be an interesting study. Another possibility is testing more models, or different versions of the same model against itself, to see if there is improvement over time, or if a paid version is worth buying instead of using the free model.

VIII. Acknowledgments

Much appreciation goes to Prof. Babafemi Sorinolu for his support and guidance throughout the entire project, from the initial stages all the way through the writing of this paper. Furthermore, many thanks are given to the human evaluators for their willingness to help and to give up their time to go through all the data.

References

[1] M. Webster, "149 AI statistics: The present and future of AI [2024 stats]," authorityhacker.com, https://www.authorityhacker.com/ai-statistics/ (accessed Nov. 29, 2024).

[2] J. Howarth, "57 New Artificial Intelligence Statistics (Nov 2024)," Exploding Topics, https://explodingtopics.com/blog/ai-statistics (accessed Nov. 29, 2024).

[3] B. Kennedy, A. Tyson, and E. Saks, "Public awareness of artificial intelligence in Everyday Activities," Pew Research Center, https://www.pewresearch.org/science/2023/02/15/public-awareness-of-artificial-intelligence-in-everyday-activities/ (accessed Nov. 29, 2024).

[4] Y. Zhang et al., "Siren's song in the AI Ocean: A survey on hallucination in large language models," arXiv.org, https://doi.org/10.48550/arXiv.2309.01219 (accessed Nov. 29, 2024).

[5] N. Gillespie, S. Lockey, C. Curtis, and J. Pool, "Trust in artificial intelligence," KPMG, https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-in-artificial-intelligence.html (accessed Nov. 29, 2024).

[6] O. Sidoti and J. Gottfried, "About 1 in 5 U.S. teens who've heard of CHATGPT have used it for schoolwork," Pew Research Center, https://www.pewresearch.org/short-reads/2023/11/16/about-1-in-5-us-teens-whove-heard-of-chatgpt-have-used-it-for-schoolwork/ (accessed Nov. 29, 2024).

[7] J. Schiel, "How high school students use and perceive technology at home and school," ACT, http://www.act.org/content/act/en/research/pdfs/R2412-How-HS-Students-Use-and-Perceive-Technology-at-Home-and-School-2024-07.html (accessed Nov. 29, 2024).

[8] "NAEP long-term trend assessment results: Reading and Mathematics," The Nation's Report Card, https://www.nationsreportcard.gov/highlights/ltt/2022/ (accessed Nov. 29, 2024).

[9] A. Auletto, "Student performance in math domains hero image," Center For Education Efficacy Excellence and Equity, https://e4.northwestern.edu/2023/05/31/math-domains/ (accessed Nov. 29, 2024).

[10] J. Ahn, R. Verma, R. Lou, D. Liu, and R. Zhang, "Large language models for mathematical reasoning: Progresses and challenges," ARXIV, https://arxiv.org/html/2402.00157v1 (accessed Nov. 29, 2024).

[11] A. Gandolfi, "GPT-4 in education: Evaluating Aptness, reliability, and loss of coherence in solving calculus problems and grading submissions - International Journal of Artificial Intelligence in Education," SpringerLink, https://link.springer.com/article/10.1007/s40593-024-00403-3 (accessed Nov. 29, 2024).

[12] H. Ramanathan and R. Palaniappan, "Comparison of three large language models as middle school math tutoring assistants.," Journal of Emerging Investigators, https://hocom.tw/Uploads/userfiles/files/jck4ahzdvja3fr8.pdf (accessed Nov. 29, 2024).

[13] C. Spreitzer, O. Straser, S. Zehetmeier, and K. Maaß, "Mathematical modelling abilities of Artificial Intelligence Tools: The case of chatgpt," MDPI, https://doi.org/10.3390/educsci14070698 (accessed Nov. 29, 2024).

[14] J. He-Yueya, G. Poesia, R. E. Wang, and N. D. Goodman, "Solving math word problems by combining language models with symbolic solvers," arXiv.org, https://doi.org/10.48550/arXiv.2304.09102 (accessed Dec. 10, 2024).

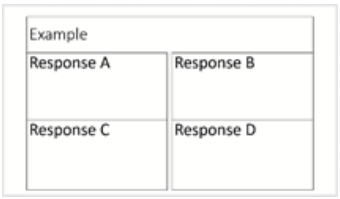

Fig. 9. Locations of Answers

IX. Appendix A

See below the statement given to the evaluators, giving them specific instructions on how the answers were organized and specific definitions of the clarity and valuability criteria.

Evaluator Instructions

Below you will find more detailed instructions, criteria, and examples to help you understand your role in the project. Please reach out if you have any further questions!

General Instructions

Each slide in the attached PowerPoint presentation corresponds to a different question. Read the full question, answer, and each of the four responses. All questions and answers were taken from verified online, open access textbooks.

Note the location of the boxes and the letter associated: [see 9]

Though the same four AI Models answered each question, their response locations have been randomized for each question to avoid any bias associated with a preferred model or a particular position.

All text in the responses is directly copied from the AI Models and the only adjustments made were to make any equations more readable, mostly by using the equation generator. For example if the model wrote 1/4, I adjusted it to say 1/4 instead to align better with what you would see on screen on the actual website of the model.

There are 3 different mathematical domains being tested: Algebra, Calculus, and Statistics, with 10 questions in each. The first 15 questions are simpler, single answer questions that would be found in a multiple choice section, for example. The second 15 are more complex, with additional explanation required or multiple parts for each question. A note about some of the statistics questions: When solving for the p-value, no additional tables were given, so some of the models did not compute an answer for that part and only explained the steps to get the answer, while others were able to calculate the answer without one. Keep that in mind while evaluating those responses.

For each question, as seen in the Google Form attached, you will be asked to rate the responses against each other in 2 separate categories, explained below. Each criteria is on a scale from 1 to 4, with 1 being the worst response in that category, and 4 being the best response in that category. Please give each response a different value; meaning if you assign Response A with a "4" in Clarity for Question 1, do not also assign Response B a "4" in Clarity. Otherwise, your evaluation will be invalid for that question.

Criteria

Clarity Criteria

This category is about how well the AI model is able to communicate its response. Things you should be looking for include:

Does the response use complete sentences and thoughts?

Are there any illogical jumps in mathematical thinking?

Does the response explain all of the variables used and use variables consistently?

Is there typos or random characters with no meaning? (* for emphasis is okay)

Also consider how the model gives information. It may explain much more information than needed and overcomplicate the response, or not give enough and leave you confused. For example, if the model gives a formula but doesn't explain how it applies, it should receive a lower score than a model that explains the variables and shows the substitution.

Clarity is really just about being able to understand what the model is trying to do in each question. If it follows a logical path and gives easy to read explanations, it will score higher than a model that leaves thoughts incomplete or jumps ahead in thinking without any explanation.

Valuability Criteria

This category is about how valuable the model is in teaching you how to solve the problem. So in contrast to Clarity, it may be more valuable to have a lot of steps explaining the work and the solution than a response that rushes through to the answers. Some things to consider are:

Did the response include separate steps on how to solve the problem?

Did the response give and explain an applicable formula for the question?

Did the response end with the correct solution?

Did the response help you understand how to solve the problem?

So while a response may not be as clear or concise as another response, it may be a more valuable explanation of the problem. So be careful to not assume a problem that scores high in clarity also scores high in valuability (though it definitely could!) Valuability is really about which response you would prefer if you were asking an AI model to help you solve and understand that question.