Vol.6 No.2 (2023)

Journal of Applied Learning & Teaching

ISSN : 2591-801X

Content Available at : http://journals.sfu.ca/jalt/index.php/jalt/index

Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics

David Santandreu Calonge[A] A *Department of Academic Development, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates*

Linda Smail[B] B *Associate Professor, College of Interdisciplinary Studies, Zayed University, Dubai, United Arab Emirates*

Firuz Kamalov[C] C *Associate Professor, School of Engineering, Applied Science and Technology, Canadian University Dubai, United Arab Emirates*

DOI: https://doi.org/10.37074/jalt.2023.6.2.22

Abstract

This article presents a comparative analysis of four AI chatbots with potential utilization in the fields of mathematics education and statistics, namely ChatGPT, GPT-4, Bard, and LLaMA. Our objective is to evaluate and compare the features, functionalities, and potential applications of these platforms within the domains of calculus and statistics. By examining their strengths and limitations, this study aims to provide insights into the selection and implementation of AI chatbots in calculus and statistics to enhance student learning. The results of the comparative analysis reveal that, while not perfect, GPT-4 outperforms ChatGPT, Bard, and LLaMA as a learning tool in calculus and statistics. Findings also reveal that chatbots may have a positive transformational impact on higher education.

Keywords: AI chatbots; Bard; calculus; ChatGPT; comparative analysis; GPT-4; Large Language Models (LLMs); LlaMA; statistics; student learning.

Introduction

Calculus and statistics are vital subjects that require effective teaching and learning methods to enhance students' engagement and comprehension. With the advancements in artificial intelligence (AI) and natural language processing (NLP), AI chatbots have emerged as promising tools for supporting students in higher education.

Kuhail et al. (2023) argued that chatbots provide a "cost-effective solution" (p. 2) to personalize learning activities, support educators, and "develop deep insight into learners' behaviour" (p. 1). AI chatbot platforms have gained significance in higher education (Singh Gill et al., 2023; Sok & Heng, 2023; Rudolph et al., 2023a; Tlili et al., 2023; Okonkwo & Ade-Ibijola, 2021; Hwang & Chang, 2021; Sandu & Gide, 2019). Moreover, the literature suggests that chatbots have the potential to enhance students' learning experience (s) in mathematics education (Castevecchi, 2023; Wardat et al., 2023) and statistics (Lee & Yeo, 2022), offering innovative solutions for learning, problem-solving, and concept clarification. They can provide personalized support, immediate feedback, interactive problem-solving, and adaptive instruction, fostering engagement and improving learning outcomes.

While there exist several studies that consider the performance of AI chatbots in mathematics problem solving, they are limited in two ways: (i) no notable analysis of Bard and LLaMA, and (ii) no analysis in statistics. This article fills the gap in the literature by evaluating and comparing four popular AI chatbot platforms, namely ChatGPT (GPT-3.5), GPT-4, Bard, and LLaMA 13-B, with a focus on their applicability and potential benefits in calculus and statistics. By examining their unique features and applications, this study aims to assist students (and educators) in selecting appropriate AI chatbot platforms to enhance their learning (and teaching) experience(s) in calculus and statistics.

Background

Benefits of using chatbots in higher education

There are several potential benefits to using chatbots in higher education (Kamalov et al., 2023). One of the main benefits is the ability to provide students with access to personalized and on-demand learning support. With chatbots, students can ask questions and receive immediate constructive feedback, which can help to reduce the workload on educators and improve the overall learning experience for students.

Another benefit of using chatbots is the ability to scale educational services (Neumann et al., 2021). Chatbots can handle large numbers of student inquiries simultaneously, which can be particularly useful in large classroom settings or in situations where there is a high demand for

educational support. This can also help decrease educators' workload and ensure that all students have access to the (individual) support they need to succeed. Findings from a study by Chen et al. (2023) revealed that chatbots had tremendous potential to help students "learn basic content in a responsive, interactive, and confidential way" (p. 1).

Additionally, chatbots have the potential to improve the efficiency of educational delivery (Huang et al., 2022). Educators can create customized learning pathways for students, which can help to ensure that students are receiving the most relevant and effective support. This can help to improve student outcomes and reduce the overall time and resources required to complete a course of study.

Drawbacks and challenges of using chatbots in higher education

While there are several potential benefits to using chatbots in higher education, there are also some drawbacks, limitations, and challenges (i.e., ethical (Popenici, 2023; Kamalov et al., 2021)) that need to be considered (Rasul et al., 2023; Rudolph et al., 2023b). Limna et al. (2023) argued, for instance, that chatbots such as ChatGPT had "caused immense concerns in education", particularly to those disciplines that "rely heavily on written assignments" (p. 3). One of the main drawbacks is the inability to fully replicate the experience of interacting with a human educator (Chen et al., 2023; Santandreu Calonge et al., 2023; Kamalov et al., 2023). This could lead to a loss of personal connections and a reduction in the quality of educational support.

Another challenge of using chatbots in higher education is the potential for harmful bias (Rasul et al., 2023; Kooli, 2023). AI systems can be biased if they are trained on biased data. This can lead to the amplification of existing biases and the exclusion of certain groups of students. It is important for educators to be aware of this potential issue and to take steps to mitigate it, such as by ensuring that chatbots are trained on a diverse and inclusive dataset. Therefore, continuous improvement and evaluation of the AI model are crucial.

A final challenge of using chatbots in higher education is the potential for technical issues (Yang & Evans, 2019). Chatbots rely on complex algorithms and sophisticated machine learning models, which can be prone to errors and glitches. This can disrupt (a) the learning experience for students and (b) the teaching experience for educators if used in the classroom as a learning and teaching activity, therefore reducing the effectiveness of chatbots as an educational tool.

To evaluate and compare the mathematical problem-solving abilities of Large Language Models (LLMs), we selected four: ChatGPT (GPT-3.5), GPT-4, Bard, and LLaMA 13-B. The choice of those four LLMs was made to ensure diversity in the study, as each AI model has its own architecture and learning mechanisms. The selection included two LLMs that were primarily designed for generating human-like text (ChatGPT and GPT-4), one LLM designed for language-related tasks (LLaMA), and one LLM that was designed to provide detailed explanations (Bard).

We investigated the following research question: Which of the four chatbots is more accurate and less verbose for statistics and calculus prompts? Kabir et al. (2023) indicated, for instance, that 52 per cent of ChatGPT answers to 517 Stack Overflow questions were incorrect, and 77 per cent were verbose.

Pros and cons of each chatbot for helping students understand calculus and statistics

ChatGPT

ChatGPT is a chatbot developed by OpenAI that is based on a large language model. It allows the user to control the conversation in terms of length, format, level of detail, style, and language. While the main purpose of the chatbot is to simulate human conversations, it can perform a wide range of tasks, including writing computer programs, composing music, answering test questions, writing poetry, and others. ChatGPT has achieved enormous popularity within a very short period, gaining over 100 million users in less than 3 months of its initial release (Rudolph et al., 2023b).

The basic version of ChatGPT is based on the GPT-3.5 model, which is a generative pre-trained transformer developed by OpenAI. GPT-3.5 is a transformer model that is first trained on large swaths of publicly available text as a general-purpose language model. Then, the model is further fine-tuned for conversational applications using a combination of supervised and reinforcement learning methods. Since GPT-3.5 is trained on unfiltered text, it is vulnerable to bias and misinformation. In addition, ChatGPT suffers from 'hallucinations' – incorrect answers that sound plausible (Rudolph et al., 2023b).

Given its capabilities, ChatGPT has been utilized in various educational domains (Lee, 2023; Qadir, 2023; Santandreu Calonge et al., 2023; Wardat et al., 2023). Wardat et al. (2023) showed that ChatGPT has the potential to provide students with mathematical knowledge. At the same time, the authors cautioned about its weaknesses in certain topics, such as geometry. The accuracy and effectiveness of ChatGPT solutions depend on the complexity of the equation, input data, and the instructions given to the chatbot. Ellis and Slade (2023) presented ChatGPT's capabilities in statistics and data science education, providing examples of how ChatGPT could help in developing course materials. A recent survey of 110 students enrolled in a mathematics course showed that students quickly adopted the ChatGPT tool, exhibiting high confidence in their responses and general usage in the learning process, alongside a positive evaluation (Sánchez-Ruiz et al., 2023). On the other hand, the development of lateral competencies was a cause for concern.

Pros

- Wide knowledge base: ChatGPT has been trained on a diverse range of topics, including calculus and statistics so that it can provide relevant information and explanations.

- Conversational nature: Students can engage in an informal dialogue with ChatGPT, asking questions and seeking clarifications, which can enhance their understanding and interest.

- Availability: ChatGPT is readily accessible through various platforms (including smartphones), making it convenient for students to seek help anytime, anywhere.

Cons

- Limited context understanding: ChatGPT might occasionally provide incorrect, incomplete, or irrelevant information due to its inability to fully grasp the context of a specific calculus question.

- Lack of visuals: Graphical representations and visual aids are often crucial in understanding calculus and statistics concepts, which ChatGPT cannot provide directly.

GPT-4

GPT-4 is a more advanced version of the GPT-3.5 language model developed by OpenAI. GPT-4 is commercially available for users under the name ChatGPT Plus. The main difference between the two versions of GPT is the size of the models, where GPT-4 consists of a much larger number of parameters than its predecessor. Although GPT-3.5 and GPT-4 show similar performance on most routine conversation tasks, the latter achieves significantly better performance on more advanced tasks, including solving mathematics questions (OpenAI Blog, 2023). For example, GPT-4 achieved over 40% percentile on the AP Calculus exam, while GPT-3.5 achieved 0%. Recent findings by Abramski et al. (2023) show that GPT-4 produces a five-fold semantically richer, more emotionally polarized perception with fewer negative associations compared to older versions of GPT. A large-scale study based on 4,550 MIT exam questions in mathematics, computer science, and electrical engineering showed that GPT-3.5 can solve a third of the problems, while GPT-4 is able to achieve a near-perfect score (Zhang et al., 2023).

Pros

- Improved contextual understanding: GPT-4 is expected to have better contextual comprehension compared to previous models, which may result in more accurate and complete responses.

- Enhanced knowledge base: GPT-4 could be trained on an updated and larger dataset, allowing it to offer more comprehensive and up-to-date information on calculus and statistics.

- Potential for more specialized models: GPT-4's architecture might be used as a basis for domain-specific models that focus solely on calculus and statistics, providing more targeted assistance.

Cons

- Potential for errors: Although GPT-4 may have a better contextual understanding, it is still a language model and can make mistakes or generate inaccurate information.

Bard

Bard is a generative artificial intelligence chatbot developed by Google. Its current version is based on the PaLM large language model, which is a transformer-based model consisting of 520 billion parameters. Bard was released to compete with the rival ChatGPT. It garnered lukewarm reception due to initial mishaps. Unlike the GPT models, Bard has direct access to the internet. A study by Plevris et al. (2023) showed Bard performs better than ChatGPT on math problems that are available online, while it underperforms on original questions. Evaluation of the mathematics performance of Bard on the mathematics test of the Vietnamese National High School Graduation Examination showed that it lagged ChatGPT (Nguyen et al., 2023). Despite the backing of Google, Bard is a relatively underutilized software with very few applications and studies in the field of education.

Pros

- Tailored for education: Bard is an AI language model specifically designed for educational purposes, including teaching subjects like calculus and statistics (Kamalov et al., 2023).

- Curriculum alignment: Bard can align its explanations and guidance with specific curricula, ensuring that students receive targeted assistance based on their educational needs.

- Pedagogical approach(es): Bard incorporates instructional strategies to enhance learning, such as providing step-by-step explanations, interactive examples, and adaptive feedback.

Cons

- Limited knowledge outside of educational content: Bard's expertise might be focused on educational topics, potentially limiting its ability to provide insights or answer questions beyond the scope of calculus and statistics.

- Dependency on available curriculum: The effectiveness of Bard heavily relies on the quality and coverage of the curriculum it is aligned with. Gaps or discrepancies in the curriculum may affect the support it offers (and the accuracy of its responses).

LLaMA

LLaMA is a large language model developed by Meta. Its developers claimed that the 13 billion parameter version of the model outperformed the much larger ChatGPT on several NLP tasks. Recently, the next-generation model

LLaMA 2 was released in partnership with Microsoft based on larger training data. Unlike other major chatbots, LLaMA is open-source software. Its relatively small size and open-source nature make it an attractive alternative to other existing chatbots. Touvron et al. (2023) showed that LLaMA is capable of outperforming Bard and ChatGPT on several NLP tasks. Similarly, Liu et al. (2023) showed that LLaMA can outperform other major chatbots in arithmetic problem-solving.

Pros

- Multimodal learning experience: LLaMA combines text-based information with visual and interactive elements, making it effective in conveying complex calculus and statistics concepts.
- Hands-on practice: LLaMA often provides interactive exercises and simulations, allowing students to actively engage with the subject matter and reinforce their understanding.
- Adaptive learning: LLaMA can adapt to the user's progress and adjust the difficulty level of the content accordingly, providing personalized learning experiences.

Cons

- Limited availability: As of the knowledge cutoff date, LLaMA is not widely accessible or integrated into various platforms, potentially limiting its reach to students.
- Resource-intensive: The integration of multimedia elements and interactive features might require robust hardware or an internet connection, which could be a barrier for some students, and in disadvantaged contexts (Shah & Calonge, 2023).

Each of these LLMs has its own advantages and limitations. Depending on the students' preferences, learning styles, and availability, they can choose the most suitable tool or combination of tools to enhance their understanding of calculus and statistics.

Methods

Seven calculus and five statistics questions were submitted to ChatGPT, GPT-4, Bard, and LLaMA 13-B via single prompts, as shown in Table 1. Each prompt was entered individually with the original question and answer choices reproduced verbatim. Each prompt was carefully designed to cover a broad range of calculus and statistical concepts. Also, the prompts varied in the level of difficulty to allow for a more in-depth analysis of the LLMs' problem-solving capabilities and to ensure a fair assessment of their mathematical skills.

Table 1. 12 prompts.

Prompts		
1	Calc	I have the below given information: "the temperature in Austin (in °F) is approximated by $T(x) = 37 \sin\left[\frac{2\pi}{365}(x-101)\right] + 25$ where T(x) is the temperature on day x, with x = 1 corresponding to Jan 1 and x = 365 corresponding to Dec 31". Using this information, please estimate the temperature, to the nearest degree Fahrenheit, on day 250. Provide me with all necessary steps.
2	Calc	Calculate the limit as x goes to zero of the ratio sin(x) over x. Explain your work.
3	Stats	On average, 3 traffic accidents per month occur at a certain intersection. What is the probability that in any given month at this intersection… (a) exactly 5 accidents will occur? (b) fewer than 3 accidents will occur? (c) at least 2 accidents will occur?
4	Calc	Find the derivative of f(x)=|x| at x=0. Explain your work.
5	Calc	Explain how to find the slope of the function f(x)=x^2+2x-1 at the point (2,3).
6	Stats	Could you explain the mean square error, give me the formula to compute it, and explain the terms involved. Provide me with an example and step by step computation of the mean square error.
7	Calc	Define a new rule for calculating Riemann sums in the following way. For each subinterval, pick the halfway point between the left endpoint and the midpoint of the subinterval. Then use the selected point to calculate the height of the corresponding rectangle. Apply the new rule to find the Riemann sum for the function f(x)=x^2+5x on the interval [2, 5] using n=10 subintervals.
8	Stats	Please explain in words the following formula and give me precise examples: b1 = Σ [(xi - x) (yi - y)] / Σ [(xi - x)2]
9	Stats	In a state that did not require varicella (chickenpox) vaccination, a boarding school experienced a prolonged outbreak of varicella among its students that began in September and continued through December. To calculate the probability or risk of illness among the students, which denominator would you use, explain your choice? Number of susceptible students at the ending of the period (i.e., June). Number of susceptible students at the midpoint of the period (late October/early November). Number of susceptible students at the beginning of the period (i.e., September). Average number of susceptible students during outbreak
10	Calc	Find the constant b so that the line y=-0.5x+b meets the graph of y^2=x^3 orthogonally. Explain your steps.
11	Stats	Help me understand how the concepts of prior, likelihood, and posterior are interrelated in Bayesian Statistics, give me an example with step-by-step explanation.
12	Calc	Find the polynomial of the smallest degree that satisfies the conditions int_1^4 p(x)= 5 and int_3^6=10. Can you suggest a general rule based on this example?

Results

This section compares the features and functionalities of each of the four AI chatbot platforms, focusing on their suitability for calculus and statistics. The evaluation of the four LLMs was based on: (1) the accuracy/the correctness of the final answer to the 12 prompts, (2) verbosity and the clarity of the explanation, and (3) the presence or absence of mathematical errors. While the correctness of the answer was assessed on a binary basis, i.e., whether the answer is correct or not, the clarity of the explanation was scored on a scale of 0 to 10, with 10 being the clearer and most comprehensive answer (see Tables 2-13). The mathematical errors were classified as either major or minor based on their potential compact on the final answer.

Accuracy

In the context of this article, chatbot accuracy is the percentage of utterances that return the correct response to the prompts, as shown graphically in figures 1- 4, below.

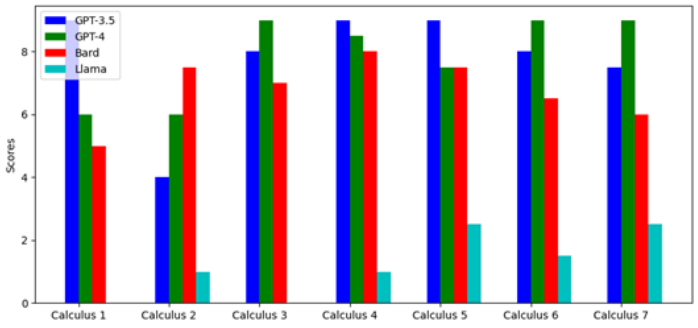


Figure 1. Accuracy scores in calculus.

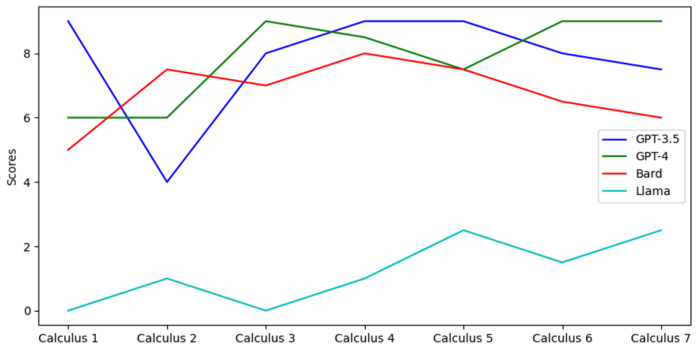


Figure 2. Calculus scores by chatbot and prompts.

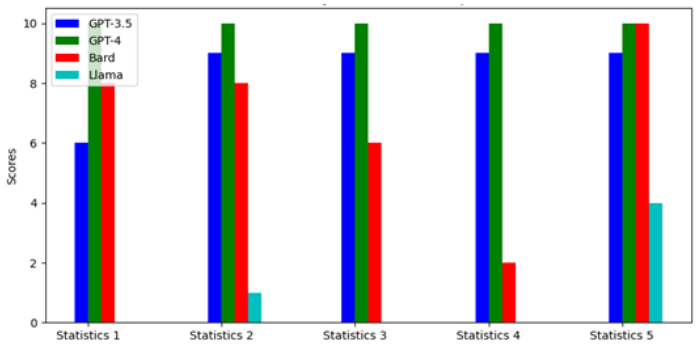


Figure 3. Accuracy scores in statistics.

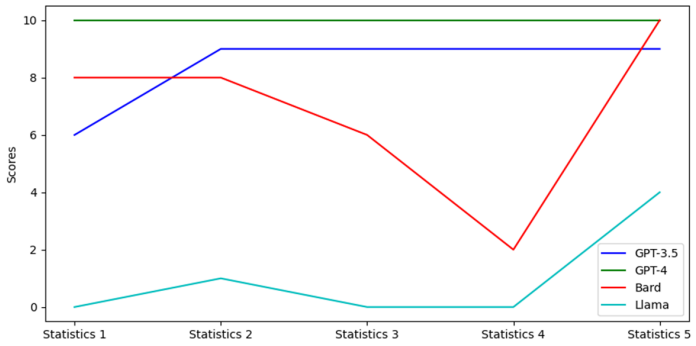


Figure 4. Statistics scores by chatbot and prompts.

Verbosity for calculus and statistics

In the context of chatbots, verbosity refers to the amount of unnecessary, irrelevant, or excessive words, phrases, or information used in the chatbot's responses (see Appendix A). A chatbot is considered verbose if it tends to provide overly detailed or convoluted answers, which can lead to a negative user experience. Zheng et al. (2023) indicated that an LLM is verbose when it "favours longer, verbose responses, even if they are not as clear, high-quality, or accurate as shorter alternatives" (Zheng et al., 2023, p. 5). Cosine similarity is a way to measure how similar two things are, e.g., two vectors or two sets of data. It calculates the cosine of the angle between the two things in a multi-dimensional space and provides a value between -1 and 1, where higher values mean greater similarity and lower values mean less similarity.

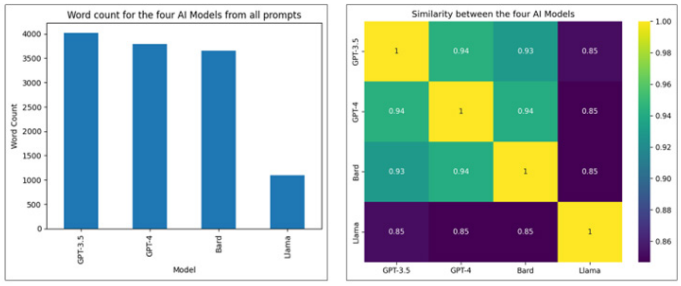


Figure 5. Verbosity (Cosine Similarity) and overlap for all 12 prompts.

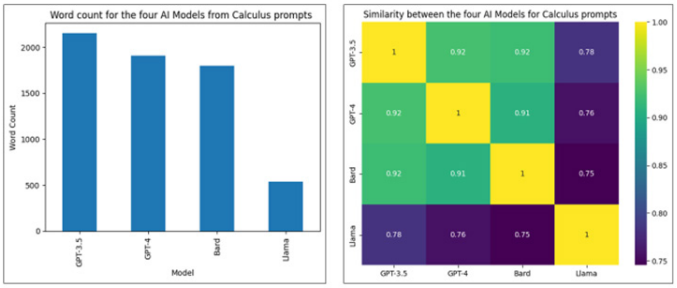


Figure 6. Verbosity (Cosine Similarity) and overlap for calculus.

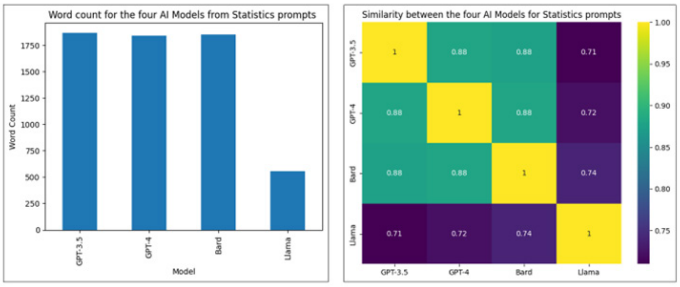


Figure 7. Verbosity (Cosine Similarity) and overlap for statistics.

Results and analysis by prompt

Prompt 1

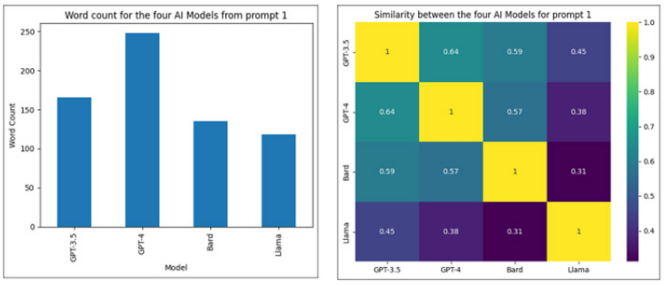


Figure 8. Verbosity (Cosine Similarity) and overlap for prompt 1.

Table 2. Answer accuracy and ratings per chatbot for prompt 1.

Chatbot	Answer accuracy	Ratings
GPT 3.5	GPT 3.5 provided an answer that is almost correct (rounding) It provided all correct steps from plugging the x value into the formula correctly, performs all necessary calculations, and comes up with an appropriate result.	▪ Final answer correct: Almost ▪ Explanation clear (0-10): 9 ▪ Mathematical mistakes: minor (rounding)
GPT-4	GPT-4 provided an incorrect answer, even though it started with a well explanation of the function, incorrectly evaluates the sine function resulting in a negative temperature.	▪ Final answer correct: No ▪ Explanation clear (0-10): 6 ▪ Mathematical mistakes: major
BARD	Bard provided accurate steps but not enough. The provided answer was not correct, there were problems with computing the angle.	▪ Final answer correct: No ▪ Explanation clear (0-10): 5 ▪ Mathematical mistakes: minor
LLaMA	Llama was completely off by using trapezoidal rule, it tried to compute the integral of the function between x=1 and x=250, which is not the right approach for this problem. It was not able to continue the answer till the end.	▪ Final answer correct: No ▪ Explanation clear (0-10): 0 ▪ Mathematical mistakes: major

Prompt 2

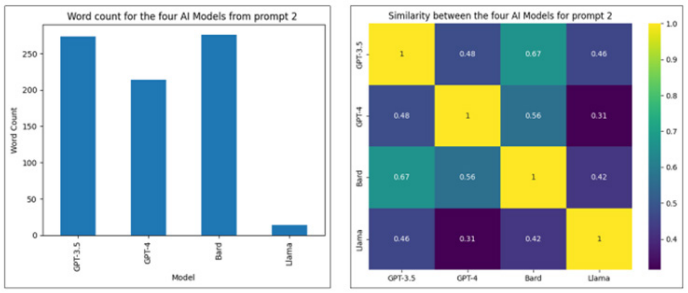


Figure 9. Verbosity (Cosine Similarity) and overlap for prompt 2.

Table 3. Answer accuracy and ratings per Chatbot for prompt 2.

Chatbot	Answer accuracy	Ratings
GPT 3.5	On the surface it appears that GPT-3.5 provides a legitimate mathematical explanation to the given question. However, a careful reading reveals several mistakes and inconsistencies in the response. The chatbot does correctly identify the squeeze theorem as a useful approach to solving the problem. But it fails to apply the theorem properly. This response provides an excellent insight into the mechanism of the chatbot which is simply trying to guess the next most likely word in the text, based on all the information that was fed into the algorithm during the training, without any logic behind it.	▪ Final answer correct: No ▪ Explanation clear (0-10): 4 ▪ Mathematical mistakes: major
GPT-4	The response provided by GPT-4 is incomplete and leaves one feeling for more information. Perhaps a follow up question to delve into more details would help in this case. The chatbot provides a short discussion of the problem without giving any technical details. Note that it makes a mistake in claiming that "This limit does not yield an indeterminate form."	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 6 ▪ Mathematical mistakes: minor
BARD	The response by Bard is arguably better than GPT. The chatbot provides 3 valid approaches to solving the problem. However, each approach either contains a mistake or incomplete. Direct substitution jumps to the conclusion using very weak logic. Squeeze theorem is applied correctly but uses prior knowledge that sin(x)<x. Taylor series is probably the best approach but there a couple of gaps in the response.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 7.5 ▪ Mathematical mistakes: minor
LLaMA	Provides a useless response.	▪ Final answer correct: No ▪ Explanation clear (0-10): 1 ▪ Mathematical mistakes: major

Prompt 3

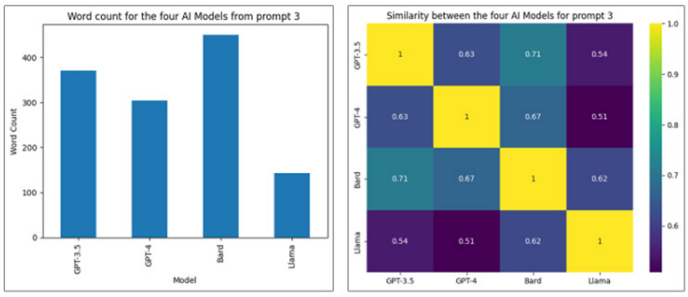


Figure 10. Verbosity (Cosine Similarity) and overlap for prompt 3.

Table 4. Answer accuracy and ratings per Chatbot for prompt 3.

Chatbot	Answer accuracy	Ratings
GPT 3.5	The explanation provided by GPT-3.5 is clear and complete, it referred to the Poisson distribution and gave detailed explanation. However, the answer was not given in full, translation to the question parts into formulas were almost correct, the part b was wrong as is translated x fewer than three to only the two cases 0 and 1.	▪ Final answer correct: No (Almost) ▪ Explanation clear (0-10): 6 ▪ Mathematical mistakes: major
GPT-4	The explanation provided by GPT-4 is clear and complete, it referred to the Poisson distribution and gave detailed explanation about what probabilities to compute, however not all probabilities were correct. Part a was wrong.	▪ Final answer correct: No (Almost) ▪ Explanation clear (0-10): 10 ▪ Mathematical mistakes: none
BARD	Bard explanation was also correct and clear and has correctly employed the Poisson distribution. However, the answers given were not correct.	▪ Final answer correct: No ▪ Explanation clear (0-10): 8 ▪ Mathematical mistakes: major
LLaMA	Llama provided an answer that is not accurate. It didn't use the Poisson distribution which is the appropriate model for these calculations. Even the provided explanation was not clear and contained major mathematical errors.	▪ Final answer correct: No ▪ Explanation clear (0-10): 0 ▪ Mathematical mistakes: major

Prompt 4

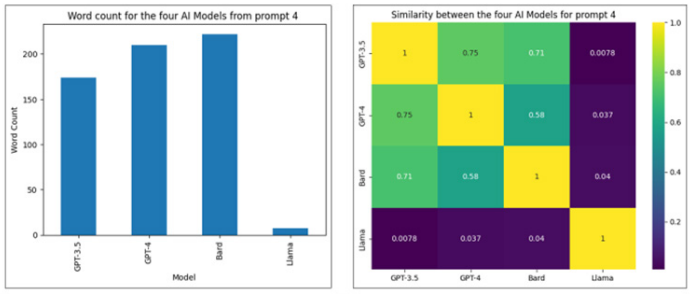


Figure 11. Verbosity (Cosine Similarity) and overlap for prompt 4.

Table 5. Answer accuracy and ratings per Chatbot for prompt 4.

Chatbot	Answer accuracy	Ratings
GPT 3.5	Overall, the chatbot provided a good response. Only a small issue with the claim about discontinuity in the end.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 8 ▪ Mathematical mistakes: minor
GPT-4	The chatbot provided a good response with some technical details which may of interest to someone who is looking for a more in-depth analysis.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 9 ▪ Mathematical mistakes: No
BARD	The response is slightly convoluted. The chatbot solves the problem using the limit definition of the derivative but goes too much into the technical details which makes it harder to follow especially given the math notation. The chatbot does make a significant mistake in the beginning to claim that the function is not continuous at x=0.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 7 ▪ Mathematical mistakes: minor
LLaMA	Most succinct response ever.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 6 ▪ Mathematical mistakes: No

Prompt 5

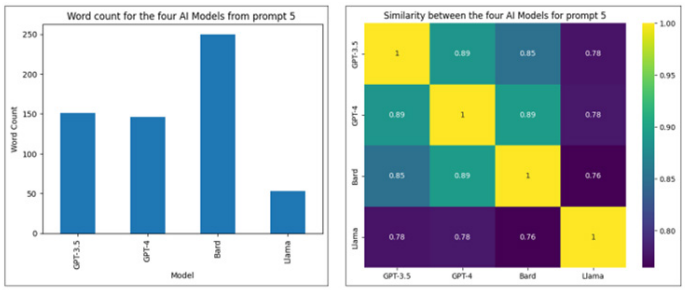


Figure 12. Verbosity (Cosine Similarity) and overlap for prompt 5.

Table 6. Answer accuracy and ratings per Chatbot for prompt 5.

Chatbot	Answer accuracy	Ratings
GPT 3.5	Overall, a good answer. I like how it breaks it down into steps so it's easier to follow. The information in step 1 could be made more concise.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 9 ▪ Mathematical mistakes: No
GPT-4	Good answer. But for a standard calculus question the response was too verbose.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 8.5 ▪ Mathematical mistakes: No
BARD	This is a standard question in calculus. The answer and explanation should be simple. While the chatbot does provide a simple explanation first, followed by a more detailed explanation, overall, it feels too verbose and hard to follow especially with all the math notation involved.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 8 ▪ Mathematical mistakes: No

LLaMA	Terrible response. 4+4-1=9?	▪ Final answer correct: No ▪ Explanation clear (0-10): 1 Mathematical mistakes: major

Prompt 6

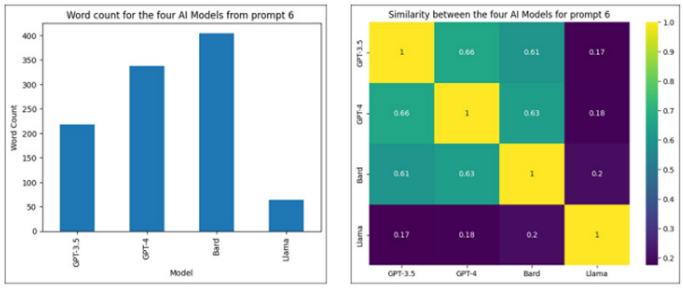


Figure 13. Verbosity (Cosine Similarity) and overlap for prompt 6.

Table 7. Answer accuracy and ratings per Chatbot for prompt 6.

Chatbot	Answer accuracy	Ratings
GPT 3.5	GPT-3.5 provided an accurate answer with a detailed explanation. Th given example was simple but answers correctly with all details.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 9 ▪ Mathematical mistakes: No
GPT-4	GPT-4 provided an even more in-depth explanation than GPT-3.5. The example was also a bit more complicated than the one provided by GOT-3.5. All calculations were correct, and all steps provided.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 10 ▪ Mathematical mistakes: No
BARD	Bard provided a clear detailed explanation and an example; however, the calculations were wrong.	▪ Final answer correct: No ▪ Explanation clear (0-10): 8 ▪ Mathematical mistakes: minor
LLaMA	Llama provided an explanation that is not clear at all and want of topic. No example was provided as requested. While the start of explanation of the formula seems accurate, the overall explanation is not as comprehensive as the other models, no mention to regression or interpretation of the MSE.	▪ Final answer correct: No ▪ Explanation clear (0-10): 1 ▪ Mathematical mistakes: major

Prompt 7

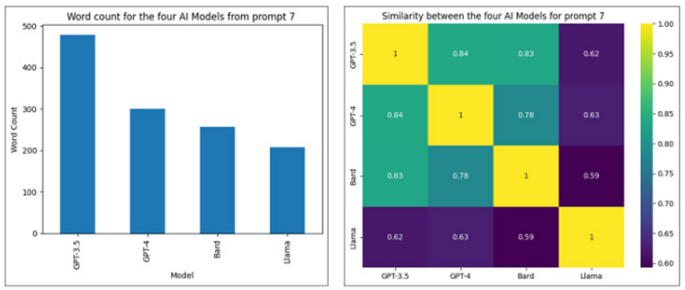


Figure 14. Verbosity (Cosine Similarity) and overlap for prompt 7.

Table 8. Answer accuracy and ratings per chatbot for prompt 7.

Chatbot	Answer accuracy	Ratings
GPT 3.5	This question is a variation of the standard Riemann sum. However, the given heuristic is not common. So, the question challenges the comprehension abilities of the chatbot. The chatbot did an excellent job of understanding the instructions and following them. It provided sufficient explanation for the reader to able to follow the process without getting confused. It correctly gave the formula for calculating the evaluation point. However, it did not follow its own formula.	▪ Final answer correct: No ▪ Explanation clear (0-10): 9 ▪ Mathematical mistakes: major
GPT-4	The chatbot did a good job understanding the instructions and providing an appropriate solution. However, the response is a bit dense and might be harder to follow for students with weaker background in math. There are 2 important flaws. First, the evaluation point is calculated incorrectly. Second, no final answer is provided. .	▪ Final answer correct: No ▪ Explanation clear (0-10): 7.5 ▪ Mathematical mistakes: major
BARD	The chatbot does a good job of understanding the instructions and provides the correct steps to calculate the answer. However, it fails to correctly calculate the evaluation point. More concrete examples of calculations would have been useful.	▪ Final answer correct: No ▪ Explanation clear (0-10): 7.5 ▪ Mathematical mistakes: major
LLaMA	Poor response.	▪ Final answer correct: No ▪ Explanation clear (0-10): 2.5 ▪ Mathematical mistakes: major

Prompt 8

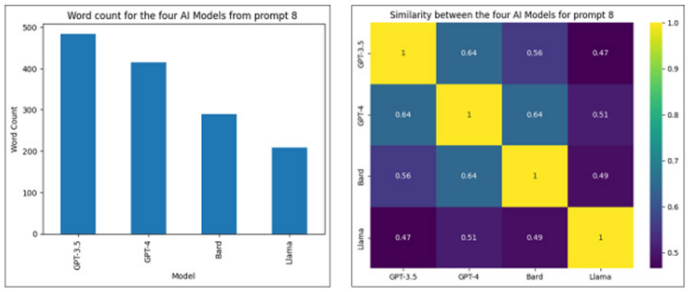


Figure 15. Verbosity (Cosine Similarity) and overlap for prompt 8.

Table 9. Answer accuracy and ratings per chatbot for prompt 8.

Chatbot	Answer accuracy	Ratings
GPT 3.5	GPT-3.5 provided accurate explanation to the question and accurate steps for the given example, however it failed to continue the steps till last step.	▪ Final answer correct: No ▪ Explanation clear (0-10): 9 ▪ Mathematical mistakes: minor
GPT-4	GPT-4 provided accurate explanation to the question and accurate computation for the given example.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 10 ▪ Mathematical mistakes: none
BARD	Bard provided an accurate explanation to the question, but not enough details about the different terms involved in the equations. The final provided answer was not correct.	▪ Final answer correct: No ▪ Explanation clear (0-10): 6 ▪ Mathematical mistakes: minor
LLaMA	Llama provided a poor response, the explanation provided does not correctly define the terms in the equation. The explanation is not clear at all as it talks about differences between pairs. A very short and simple example was provided, and steps were wrong to reach the final answer.	▪ Final answer correct: No ▪ Explanation clear (0-10): 0 ▪ Mathematical mistakes: major

Prompt 9

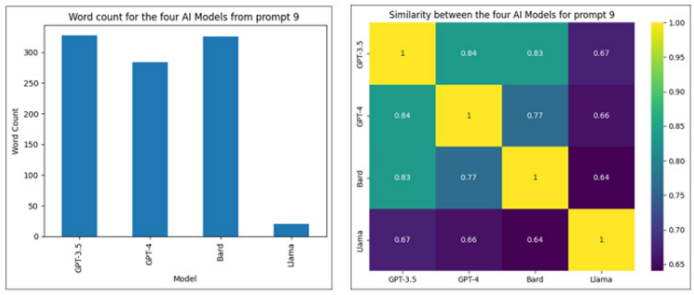
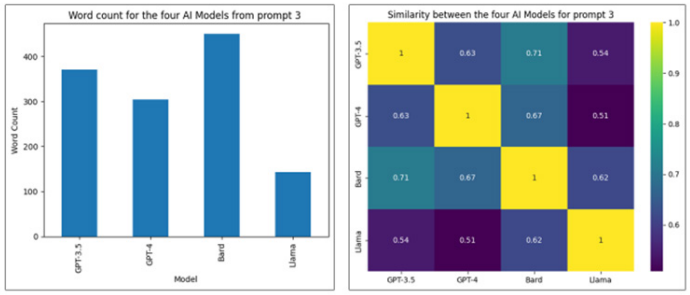


Figure 16. Verbosity (Cosine Similarity) and overlap for prompt 9.

Table 10. Answer accuracy and ratings per chatbot for prompt 9.



Prompt 10

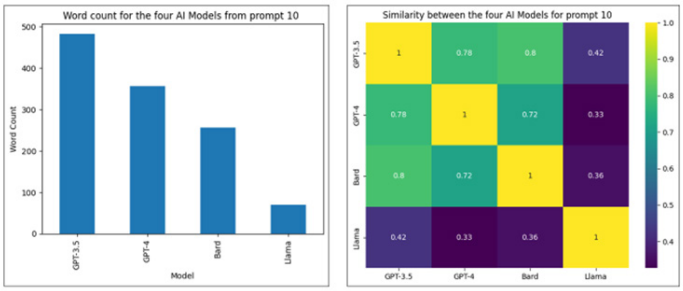


Figure 17. Verbosity (Cosine Similarity) and overlap for prompt 10.

Table 11. Answer accuracy and ratings per chatbot for prompt 10.

Chatbot	Answer accuracy	Ratings
GPT 3.5	This is a nontrivial question with a multi-step solution. Overall, it is a good attempt to solve the problem. The chatbot correctly identifies the key concepts and ingredients to solve the problem and provides a step-by-step explanation of the solution. However, GPT-3.5 fails to put it all together. It follows the correct path to solution until Step 4, after which it takes a wrong turn. While there are no major mistakes in terms of math, the chatbot pursued the wrong strategy which ultimately led to an unresolved outcome.	▪ Final answer correct: No ▪ Explanation clear (0-10): 8 ▪ Mathematical mistakes: minor
GPT-4	The chatbot provides a good step-by-step explanation of the solution. Overall, the presented solution is correct which is impressive given the level of the difficulty and the number of steps required to solve the problem. However, it makes a small mistake in a basic calculation "y = 3*(16/9) ^2 / 4 => y = 32 / 9". It is interesting to observe that while GPT-4 can tackle the problem conceptually which is the hardest part, it makes a basic calculation error especially since calculations are generally regarded as the strength of the chatbots.	▪ Final answer correct: Yes* ▪ Explanation clear (0-10): 9 ▪ Mathematical mistakes: minor *95%
BARD	The chatbot fails to recognize that the derivative is found using implicit differentiation. It also does not realize that the provided answer does not make sense.	▪ Final answer correct: No ▪ Explanation clear (0-10): 6.5 ▪ Mathematical mistakes: major
LLaMA	Nonsensical response.	▪ Final answer correct: No ▪ Explanation clear (0-10): 1.5 ▪ Mathematical mistakes: major

Prompt 11

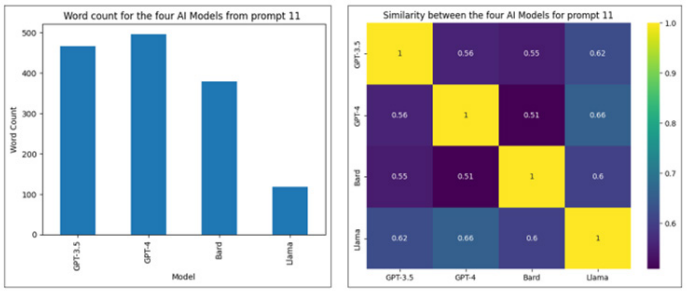


Figure 18. Verbosity (Cosine Similarity) and overlap for prompt 11.

Table 12. Answer accuracy and ratings per Chatbot for prompt 11.

Chatbot	Answer accuracy	Ratings
GPT 3.5	GPT-3.5 answer is correct and provides a clear and detailed explanation. It gave an example and explained the relationship between prior, likelihood, and posterior in Bayesian Statistics but failed in completing the answer by the end	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 9 ▪ Mathematical mistakes: none
GPT-4	GPT-4 answer is correct, it was explained in details and an example was used to provide more insights	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 10 ▪ Mathematical mistakes: none
BARD	Bard answer is correct, clearly explained with an example and enough details.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 10 ▪ Mathematical mistakes: none
LLaMA	Llama answer is correct with clear explanation, however, no example was provided, as required in the prompt.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 4 ▪ Mathematical mistakes: none

Prompt 12

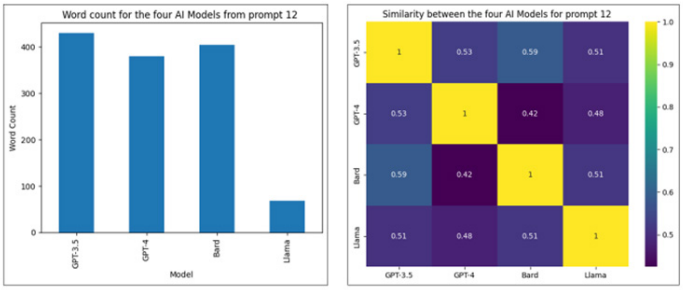


Figure 19. Verbosity (Cosine Similarity) and overlap for prompt 12.

Table 13. Answer accuracy and ratings per chatbot for prompt 12.

Chatbot	Answer accuracy	Ratings
GPT 3.5	The chatbot supplies the general strategy for solving the problem but does not execute the proposed plan. Thus, it has an incomplete understanding of the solution. While the proposed approach is far from being complete, it is presented in a clear style.	▪ Final answer correct: No ▪ Explanation clear (0-10): 7.5 ▪ Mathematical mistakes: No
GPT-4	The chatbot provides a correct and complete solution. The solution is presented clearly. However, there is a small calculation mistake "[ax^2/2 + bx] _1^4 = 5, or [(8a/2 + 4b) - (a/2 + b)]". It is puzzling that GPT-4 can solve complex problems but can stumble on a basic calculation.	▪ Final answer correct: Yes ▪ Explanation clear (0-10): 9 Mathematical mistakes: minor
BARD	The response looks legitimate on the surface, but a closer look at the details reveals multiple holes. There are serious mathematical flaws in the arguments and the final answer (deg2) is incorrect.	▪ Final answer correct: No ▪ Explanation clear (0-10): 6 ▪ Mathematical mistakes: Major
LLaMA	Incorrect solution.	▪ Final answer correct: No ▪ Explanation clear (0-10): 2.5 Mathematical mistakes: major

Discussion

Use cases in calculus and statistics

In this article, we explored potential use cases for each platform within calculus and statistics. We argue that ChatGPT and GPT-4 can be utilized in calculus and statistics to provide personalized tutoring and assistance to students. Both can generate step-by-step solutions to math problems, explain complex mathematical concepts, and offer practice exercises to reinforce learning. Students can engage in interactive conversations with ChatGPT or GPT-4 to clarify doubts, receive real-time feedback anytime, anywhere, and improve their understanding of mathematical principles.

Bard can also play a vital role in calculus and statistics. It can assist students with administrative tasks related to course registration, provide access to mathematical resources such as textbooks and study materials, and offer guidance on choosing appropriate courses for specific mathematical or statistical topics. However, it is significantly weaker than GPT-3.5 and GPT-4 in calculus and statistics. LLaMA is, unfortunately, and disappointingly not very accurate for calculus and statistics prompts.

Whilst Popenici (2023) argued that AI was facilitating the super-personalisation (p. 5) of the learning experience, Rasul et al. (2023) indicated that ChatGPT could be utilized to facilitate adaptive learning, provide personalised feedback, aid research, automate administrative services, and create innovative assessment.

Our findings indicate that chatbots can also be utilized in several ways to assist students in comprehending statistics or calculus better if they have received prior training on writing effective prompts (Eager & Brunton, 2023):

1. Concept explanation: Students, following training on prompt structuring, could engage in a conversation with a chatbot to seek explanations and clarifications on statistical or calculus concepts they find challenging. Chatbots with knowledge-tracing capabilities (Shehata et al., 2023) can provide step-by-step explanations, examples, and intuitive analogies to help students understand statistical concepts in a personalized and interactive manner.

2. Problem-solving: Students can present statistical problems or exercises to a chatbot, and it can guide them through the problem-solving process if specifically asked in the prompt. Chatbots can offer hints, ask relevant questions to trigger critical thinking, and provide guidance on the correct approach or methodology to solve the problem. It can therefore help "increase student engagement and satisfaction by relieving university staff of routine tasks and allowing them to focus on higher-order skills and mentoring" (Firat, 2023, p. 61).

3. Data analysis assistance: Students can seek help from chatbots in analyzing data sets, confirming research by Carlander-Reuterfelt et al. (2020). They can input their data, and chatbots can guide them through the appropriate statistical techniques, such as calculating measures of central tendency, conducting hypothesis tests, or creating visualizations. Chatbots can provide insights into data interpretation and explain the implications of the statistical results.

4. Real-world applications: Chatbots can showcase authentic applications of statistics or calculus to students. By discussing examples and case studies from various fields, such as social sciences, healthcare, economics, or sports, chatbots can illustrate how statistical or calculus concepts can be utilized in practical situations. Hultberg et al. (2018) argued that "making a link between often abstract concepts and pertinent examples" can help "students understand difficult ideas, thus making it easier to remember" (p. 35). This can help students grasp the relevance and significance of statistics and calculus in different domains.

5. Practice and assessment: In line with the recent extant literature in a range of disciplines, chatbots can offer interactive practice sessions and quizzes to assess students' understanding of statistical or calculus concepts. They can provide instant feedback on their answers, explain any mistakes, miscalculations, or misconceptions, and suggest further study materials or resources for improvement (Mogavi et al., 2023).

Last and not least, chatbots can serve as tireless, mobile, interactive, and personalized learning companions, offering explanations, guidance, and practice opportunities 24/7 to help students grasp statistical or calculus concepts more effectively. Its conversational nature allows for an engaging and interactive learning experience and can cater to students' individual learning styles, preferences, and needs.

Summary

In summary, the four AI chatbot platforms have a wide range of use cases in calculus and statistics, including personalized tutoring, administrative support, adaptive assessments, collaborative learning, and concept clarification. Their capabilities vary greatly in terms of responses (from very accurate to not-so-good), allowing educators and students to choose the platform that best aligns with their specific needs and goals in calculus and statistics education.

Limitations

While this study marked a crucial step in understanding the potential and limitations of LLMs in teaching calculus and statistics, it has several limitations. First, the study's focus is limited to only these two areas, which restricts the generalization of the findings to other academic disciplines. Second, the choice of the four LLMs, though considered the most well-known and used, is not exhaustive, leaving numerous other LLMs, such as Claude, Upstage, Falcon or Vicuna, unexplored. Third, the assessment of the quality of the LLMs' explanations is subjective and could differ based on individual perspectives. It is also important to bear in mind the possible bias in the chatbots' responses. Fourth, due to practical constraints, this paper could not capture the dynamic learning and evolution of the four AI models over time.

Future directions

The findings of our study indicate areas where future research on LLMs' development could focus, particularly in terms of contextual understanding and the ability to provide clear, concise, and accurate explanations of calculus and statistical prompts. We suggest training AI models using specific educational resources or textbooks commonly used in calculus and statistics, enhancing their alignment with the curriculum and their ability to provide targeted assistance. Integration with platforms such as https://www.snapxam.com/ may also improve responses' accuracy. Another suggestion for future research is to investigate the impact of using LLMs on students' performance, motivations, and self-efficacy when used along with traditional teaching methods.

Conclusion and implications

This comparative analysis provides valuable insights into the features and applications of AI chatbot platforms— ChatGPT, GPT-4, Bard, and LLaMA 13-B —in the context of calculus and statistics. Each platform offers unique functionalities

that can empower students (Hutson & Plate, 2023), enhance learning, authentic assessment (Ifelebuegu, 2023), problem-solving, and engagement in these disciplines. Wu and Yu (2023) indicated that chatbots may help improve students' learning outcomes.

Overall, chatbots have the potential to transform the way in which higher education is delivered in the classroom and online. They offer a range of benefits, including personalized and on-demand learning support, the ability to scale educational services, and improved efficiency in educational delivery. However, there are also some drawbacks and challenges that need to be considered, including the potential for academic dishonesty, plagiarism (Chaka, 2023) and cheating, privacy issues, bias, and the risk of technical issues. The findings reported here shed new light on the use of AI and LLMs in teaching and learning. Students can use this information to select an LLM that best suits their needs and complements their learning style. By carefully considering the pros and cons of using chatbots in higher education, educators can make informed decisions about whether and how to incorporate this technology into their teaching practices. Despite its limitations, the findings from this study make several contributions to the current literature and lay the groundwork for future research into the use of chatbots to improve learning and teaching in a range of academic disciplines.

Data availability statement: The datasets used/analyzed during the current study are available from the corresponding author upon reasonable request.

References

Abramski, K., Citraro, S., Lombardi, L., Rossetti, G., & Stella, M. (2023). Cognitive network science reveals bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring math anxiety in high-school students. *Big Data and Cognitive Computing, 7*(3), 124 https://doi.org/10.3390/bdcc703012424

Carlander-Reuterfelt, D., Carrera, Á., Iglesias, C. A., Araque, Ó., Rada, J. F. S., & Muñoz, S. (2020). JAICOB: A data science chatbot. *IEEE Access, 8*, 180672-180680. https://doi.org/10.1109/ACCESS.2020.3024795.

Castelvecchi, D. (2023). How will AI change mathematics? Rise of chatbots highlights discussion. *Nature, 615*(7950), 15-16. https://doi.org/10.1038/d41586-023-00487-2

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching, 6*(2), 1-11. Advanced online publication. https://doi.org/10.37074/jalt.2023.6.2.12

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers, 25*(1), 161-182. https://doi.org/10.1007/s10796-022-10291-4

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?*. arXiv preprint arXiv:2307.09009.

Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-Augmented teaching and learning practice. *Journal of University Teaching & Learning Practice, 20*(5), 02.

Ellis, A. R., & Slade, E. (2023). A new era of learning: Considerations for ChatGPT as a tool to enhance statistics and data science education. *Journal of Statistics and Data Science Education,* 1-10. https://doi.org/10.1080/26939169.2023.2223609

Firat, M. (2023). What ChatGPT means for universities: Perceptions of scholars and students. *Journal of Applied Learning and Teaching, 6*(1), 57-63. https://doi.org/10.37074/jalt.2023.6.1.22

Hultberg, P., Calonge, D. S., & Lee, A. E. S. (2018). Promoting long-lasting learning through instructional design. *Journal of the Scholarship of Teaching and Learning, 18*(3), 26-43. https://doi.org/10.14434/josotl.v18i3.23179

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning, 38*(1), 237-257. https://doi.org/10.1111/jcal.12610

Hutson, J., & Plate, D. (2023). *Human-AI collaboration for smart education: Reframing applied learning to support metacognition.* IntechOpen.

Hwang, G. J., & Chang, C. Y. (2021). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments,* 1-14. https://doi.org/10.1080/10494820.2021.1952615

Ifelebuegu, A. (2023). Rethinking online assessment strategies: Authenticity versus AI chatbot intervention. *Journal of Applied Learning and Teaching, 6*(2), 1-8. Advanced online publication. https://doi.org/10.37074/jalt.2023.6.2.2

Kabir, S., Udo-Imeh, D. N., Kou, B., & Zhang, T. (2023). *Who answers it better? An in-depth analysis of ChatGPT and Stack overflow answers to software engineering questions.* arXiv preprint arXiv:2308.02312.

Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New era of Artificial Intelligence in education: Towards a sustainable multifaceted revolution. *Sustainability, 15*(16), 12451. https://doi.org/10.3390/su151612451

Kamalov, F., Sulieman, H., & Santandreu Calonge, D. (2021). Machine learning based approach to exam cheating detection. *Plos One, 16*(8), e0254340. https://doi.org/10.1371/journal.pone.0254340

Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability, 15*(7), 5614. https://doi.org/10.3390/su15075614

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies, 28*(1), 973-1018. https://doi.org/10.1007/s10639-022-11177-3

Lee, D., & Yeo, S. (2022). Developing an AI-based chatbot for practicing responsive teaching in mathematics. *Computers & Education, 191*, 104646. https://doi.org/10.1016/j.compedu.2022.104646

Lee, H. (2023). The rise of ChatGPT: Exploring its potential in medical education. *Anatomical Sciences Education*. https://doi.org/10.1002/ase.2270

Limna, P., Kraiwanit, T., Jangjarat, K., Klayklung, P., & Chocksathaporn, P. (2023). The use of ChatGPT in the digital era: Perspectives on chatbot implementation. *Journal of Applied Learning and Teaching, 6*(1), 64-74. https://doi.org/10.37074/jalt.2023.6.1.32

Liu, T., & Low, B. K. H. (2023). *Goat: fine-tuned LLaMA outperforms GPT-4 on arithmetic tasks*. arXiv preprint arXiv:2305.14201.

Mogavi, R. H., Deng, C., Kim, J. J., Zhou, P., Kwon, Y. D., Metwally, A. H. S., ... & Hui, P. (2023). *Exploring user perspectives on chatgpt: Applications, perceptions, and implications for AI-integrated education*. arXiv preprint arXiv:2305.13114.

Neumann, A. T., Arndt, T., Kobis, L., Meissner, R., Martin, A., de Lange, P., Pengel, N., Klamma, R., & Wollersheim, H. W. (2021). Chatbots as a tool to scale mentoring processes: Individually supporting self-study in higher education. *Frontiers in Artificial Intelligence, 4,* 668220. https://doi.org/10.3389/frai.2021.668220

Nguyen, P., Nguyen, P., Bruneau, P., Cao, L., Wang, J., & Truong, H. (2023). *Evaluation of mathematics performance of Google Bard on the mathematics test of the Vietnamese national high school graduation examination.*

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence, 2*, 100033. https://doi.org/10.1016/j.caeai.2021.100033

Plevris, V., Papazafeiropoulos, G., & Rios, A. J. (2023). *Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard*. arXiv preprint arXiv:2305.18618.

Popenici, S. (2023). The critique of AI as a foundation for judicious use in higher education. *Journal of Applied Learning and Teaching, 6*(2), 1-7. https://doi.org/10.37074/jalt.2023.6.2.4

Qadir, J. (2023, May). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1-9). IEEE. https://doi.org/10.1109/EDUCON54358.2023.10125121.

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., ... & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching, 6*(1), 41-56. https://doi.org/10.37074/jalt.2023.6.1.29

Rudolph, J., Tan, S., & Tan, S. (2023a). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching, 6*(1), 364-389. https://doi.org/10.37074/jalt.2023.6.1.23

Rudolph, J., Tan, S., & Tan, S. (2023b). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching, 6*(1), 342-363. https://doi.org/10.37074/jalt.2023.6.1.9

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), 6039. https://doi.org/10.3390/app13106039

Sandu, N., & Gide, E. (2019, September). Adoption of AI-Chatbots to enhance student learning experience in higher education in India. In *2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET)* (pp. 1-5). IEEE. htpps://doi.org/10.1109/ITHET46829.2019.8937382

Santandreu-Calonge, D., Medina-Aguerrebere, P., Hultberg, P., & Shah, M. A. (2023). Can ChatGPT improve communication in hospitals?. *Profesional de la información, 32*(2). https://doi.org/10.3145/epi.2023.mar.19

Shah, M., & Calonge, D. S. (2023). Refugees' experiences with online higher education: Impact and implications through the pandemic. *Journal of Applied Learning and Teaching, 6*(1), 209-221. https://doi.org/10.37074/jalt.2023.6.1.21

Shehata, S., Calonge, D. S., Purnell, P., & Thompson, M. (2023, July). Enhancing video-based learning using knowledge tracing: Personalizing students' learning experience with ORBITS. In *Proceedings of the 18th workshop on innovative use of NLP for Building Educational Applications (BEA 2023)* (pp. 100-107). https://doi.org/10.18653/v1/2023.bea-1.8

Singh Gill, S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., Fuller, S., Singh, M., Arora, P., Parlikad, A. K., Stankovski, V., Abraham, A., Ghosh, S. K., Lutfiyya, H., Kanhere, S. S., Bahsoon, R., Rana, O., Dustdar, S., Sakellariou, R., … Buyya, R. (2023). *Transformative effects of ChatGPT on modern education: Emerging era of AI chatbots*. arXiv e-prints, arXiv-2306. https://doi.org/10.1016/j.iotcps.2023.06.002

Sok, S., & Heng, K. (2023). *ChatGPT for education and research: A review of benefits and risks*. SSRN 4378735. http://dx.doi.org/10.2139/ssrn.4378735.

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*(1), 15. https://doi.org/10.1186/s40561-023-00237-x

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). *Llama: Open and efficient foundation language models*. arXiv preprint arXiv:2302.13971.

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(7), em2286. https://doi.org/10.29333/ejmste/13272

Yang, S., & Evans, C. (2019, November). Opportunities and challenges in using AI chatbots in higher education. In *Proceedings of the 2019 3rd international conference on education and e-learning* (pp. 79-83). https://doi.org/10.1145/3371647.3371659

Wu, R., & Yu, Z. (2023). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13334

Zhang, S. J., Florin, S., Lee, A. N., Niknafs, E., Marginean, A., Wang, A., ... & Drori, I. (2023). *Exploring the MIT mathematics and EECS curriculum using Large Language Models*. arXiv preprint arXiv:2306.08997.

Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-judge with MT-Bench and Chatbot Arena*. arXiv preprint arXiv:2306.05685.