MERGING LANES: WHERE E-LEARNING DIVERSITY MEETS FUTURE TRENDS

VOLUME 11, 2026
e-ISBN : 978-629-98755-9-8

eISBN 978-629-98755-9-8
9 786299 875598

SIG CS@e-Learning
Unit Penerbitan
Jabatan Sains Komputer & Matematik
Kolej Pengajian Pengkomputeran, Informatik & Matematik
Universiti Teknologi MARA Cawangan Pulau Pinang

MERGING LANES: WHERE E-LEARNING DIVERSITY MEETS FUTURE TRENDS

Advisor
Dr. Nor Hanim Abd Rahman,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

Chief Editor
Ts. Jamal Othman,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

Editors
Ts. Dr. Syarifah Adilah Mohamed Yusoff,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

Dr Arifah Fasha Rosmani,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

Mohd Saifulnizam Abu Bakar,
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

Published by:
Unit Penerbitan Jabatan Sains Komputer & Matematik (JSKM)
Bahagian Hal Ehwal Akademik (BHEA)
Universiti Teknologi MARA
Cawangan Pulau Pinang
13500 Permatang Pauh
Pulau Pinang
Malaysia

e ISBN : 978-629-98755-9-8

e-ISBN : 978-629-98755-9-8

SIG : e-Learning@CS
https://appspenang.uitm.edu.my/sigcs/
Publication Date : 15 – Apr - 2026

A STATISTICAL INVESTIGATION OF AI TOOLS' ACCURACY IN SOLVING ALGEBRA, CALCULUS, AND STATISTICS PROBLEMS: COMPARATIVE ANALYSIS OF CHATGPT AND GEMINI

*Norazah Umar[1], Nurhafizah Ahmad[2], Jamal Othman[3], Muniroh Hamat[4]
[*] norazah191@uitm.edu.my[1], nurha9129@uitm.edu.my[2] , jamalothman@uitm.edu.my[3], muniroh@uitm.edu.my[4]

[1,2,3,4]Jabatan Sains Komputer & Matematik (JSKM),
Universiti Teknologi MARA Cawangan Pulau Pinang, Malaysia

** Corresponding author*

ABSTRACT

This study presents a descriptive comparative investigation of the mathematical accuracy of two widely used large language model (LLM) tools, ChatGPT and Gemini, across three core domains: Algebra, Calculus, and Statistics. The increasing adoption of generative AI in higher education has raised concerns about the reliability of AI-generated mathematical solutions, particularly when outputs appear coherent but contain hidden reasoning gaps. To examine domain-specific performance, both tools were tested using an identical prompt protocol, and only first responses were recorded to reflect typical student usage. Accuracy was evaluated using final-answer correctness and summarized using descriptive statistics, reported as percentage of correct solutions by domain. Results indicate that both tools achieved consistently high accuracy across all domains, exceeding 88%. ChatGPT demonstrated higher accuracy in Algebra (97.22%) compared to Gemini (91.67%), suggesting stronger performance on symbolic manipulation and structured equation-based tasks. In contrast, Gemini achieved perfect accuracy in both Calculus and Statistics (100% each), outperforming ChatGPT in those domains (88.89% and 94.44%, respectively). These findings indicate that LLM effectiveness in mathematics is domain-dependent rather than uniform, with each system exhibiting distinct strengths. Overall, the study suggests that AI tools can serve as useful computational assistants in mathematics learning and practice, but domain sensitivity implies that outputs should be interpreted cautiously and verified, especially in formal assessment contexts. Future work should expand the problem set, incorporate step-validity scoring, and evaluate performance under reworded and out-of-distribution problem conditions to better assess reasoning robustness.

Keywords : ChatGPT, Gemini, AI, Large Language Models, Mathematical Accuracy.

1.0 Introduction

The rise of Artificial Intelligence (AI) has revolutionized many fields, including education, where AI-driven systems such as ChatGPT and Gemini are becoming essential tools for students and educators alike. With the increasing integration of these systems into learning environments, it is crucial to evaluate their mathematical accuracy and reasoning capabilities. This study aims to provide a comparative analysis of two leading AI models ChatGPT and Gemini across three mathematical domains: Algebra, Calculus, and Statistics.

The purpose of this research is to assess the mathematical performance of these AI systems by evaluating their accuracy in solving university-level mathematics questions. By comparing the

performance of ChatGPT and Gemini, the study seeks to identify strengths and weaknesses in each model, as well as to understand how AI tools perform across different mathematical domains. Given that mathematics involves both procedural tasks and reasoning-based problem-solving, examining how these AI systems handle multi-step reasoning challenges is particularly important.

This paper is structured as follows: Section 3 outlines the methodology employed in this study, Section 4 presents the results and analysis, and Section 5 discusses the implications of the findings for educational use. Through this comparative analysis, the study contributes to a growing body of literature on the efficacy of AI in mathematics education and aims to provide insight into the limitations and potential applications of these models

2.0 Literature Review

The integration of AI in education has grown significantly over the past decade, particularly with the introduction of large language models (LLMs) such as OpenAI's GPT series and Google DeepMind's Gemini models. These models are trained on vast datasets and are capable of performing tasks that range from simple computational problems to complex reasoning exercises. However, while these tools demonstrate impressive capabilities, their performance in mathematics remains a topic of ongoing research (Duan et al., 2025; Jahin et al., 2025).

Previous studies have highlighted the strengths and limitations of LLMs in mathematical problem-solving. For instance, Brown et al. (2020) demonstrated that GPT-based models excel in symbolic manipulation tasks, such as solving algebraic equations and performing basic arithmetic. Their ability to perform step-by-step transformations in well-structured algebraic problems is often seen as a significant advantage. On the other hand, Lin et al. (2022) cautioned that LLMs tend to struggle with tasks requiring deep logical reasoning or those that deviate from typical patterns seen during training. This limitation can lead to errors in more complex areas, such as calculus and statistics, where multi-step reasoning and formula-based application are essential.

Further research by Boye and Moëll (2025) and Edwards (2025) explored how LLMs, despite generating correct final answers, can still exhibit reasoning flaws that undermine the validity of intermediate steps. This is particularly evident in tasks that require nuanced deductive processes or reasoning through unfamiliar problem types. Given this, it becomes essential not only to evaluate the final accuracy but also to investigate whether the underlying logical steps are valid.

Gemini, a newer AI model from Google DeepMind, has been positioned as a competitor to GPT-based systems. Preliminary findings suggest that Gemini models exhibit superior performance in domains that require procedural accuracy, such as calculus and statistics (DesignTalks, 2024). The

model's ability to maintain consistency in solving integration and differentiation problems aligns with its design, which emphasizes computational proficiency and accuracy. However, questions remain regarding Gemini's handling of more complex, multi-step problems in algebraic reasoning (Jahin et al., 2025).

Given these findings, this study seeks to bridge the gap in existing literature by providing a direct comparison of ChatGPT and Gemini in terms of their mathematical accuracy across algebra, calculus, and statistics. It aims to clarify the strengths and weaknesses of each model and to explore how task-specific characteristics, such as the need for procedural versus symbolic reasoning, influence AI performance.

3.0 Methodology

This study employed a quantitative descriptive comparative design to evaluate and compare the mathematical accuracy of ChatGPT and Gemini across three domains: Algebra, Calculus, and Statistics. The methodological structure was aligned with the performance outcomes presented in Figure 1, where comparison is based on domain-level percentage accuracy. The main purpose was to describe observable performance differences between the two AI systems rather than to generalize statistically beyond the tested dataset.

3.1 Research Design

A cross-sectional task-based evaluation was conducted. Both AI tools were tested using the same set of mathematics questions under standardized conditions. This approach reflects common benchmarking strategies used in LLM evaluation, where models are compared by performance across subject categories to identify domain-specific strengths and weaknesses (Duan et al., 2025; Jahin et al., 2025). The decision to separate Algebra, Calculus, and Statistics was motivated by evidence that LLM mathematical performance is not uniform and often varies depending on task structure and reasoning requirements (Lin et al., 2022; John, 2025).

A structured set of university-level questions was developed to represent the three domains. Algebra questions emphasized symbolic manipulation and multi-step transformations, calculus questions emphasized procedural application of differentiation and integration rules, and statistics questions emphasized probability and computational statistics. The problem set intentionally included multi-step items to ensure that the tools were evaluated on tasks requiring sequential reasoning rather than only single-step computation. This decision is consistent with recent literature noting that multi-step mathematical reasoning is a common failure point for LLMs even when outputs appear fluent and confident (Boye & Moëll, 2025; Edwards, 2025).

3.2 Data Collection Procedure

Each question was entered into ChatGPT and Gemini using identical prompt wording to reduce prompt-induced bias. Only the first response produced by each tool was recorded, and no follow-up prompts, hints, or iterative corrections were provided. This procedure was chosen to mirror typical student usage, where users frequently rely on the first generated output and may not systematically verify intermediate reasoning steps (Urban et al., 2024). All AI outputs were saved verbatim and compared against verified answer keys prepared before testing.

Performance was evaluated using final-answer correctness only. Each response was coded as correct (1) or incorrect (0) based on agreement with the validated solution. For each tool, the total number of correct responses was computed within each domain and converted into percentage accuracy, which forms the basis of the comparative results displayed in Figure 1. Accuracy-percentage reporting is widely used in mathematical AI benchmarking because it provides a direct, transparent summary of observed performance across categories (Duan et al., 2025; Jahin et al., 2025).

3.3 Data Analysis

The study used descriptive statistics. Results were summarized using frequencies of correct versus incorrect responses and percentage accuracy by domain for each AI tool. Domain-level summaries were then presented visually to highlight performance differences across Algebra, Calculus, and Statistics. No inferential statistical tests were conducted because the goal was to describe tool performance within the collected dataset rather than to estimate population parameters.

This methodology was designed to capture domain-dependent performance variation, consistent with findings that LLMs may succeed in well-structured procedural tasks but show instability when reasoning demands increase or when problems deviate from learned patterns (Lin et al., 2022; John, 2025). By applying standardized prompts and reporting domain-specific accuracy descriptively, the study provides a clear comparison of how ChatGPT and Gemini perform across distinct mathematical disciplines, while remaining aligned with established LLM benchmarking approaches (Duan et al., 2025; Jahin et al., 2025).

4.0 Results & Discussion

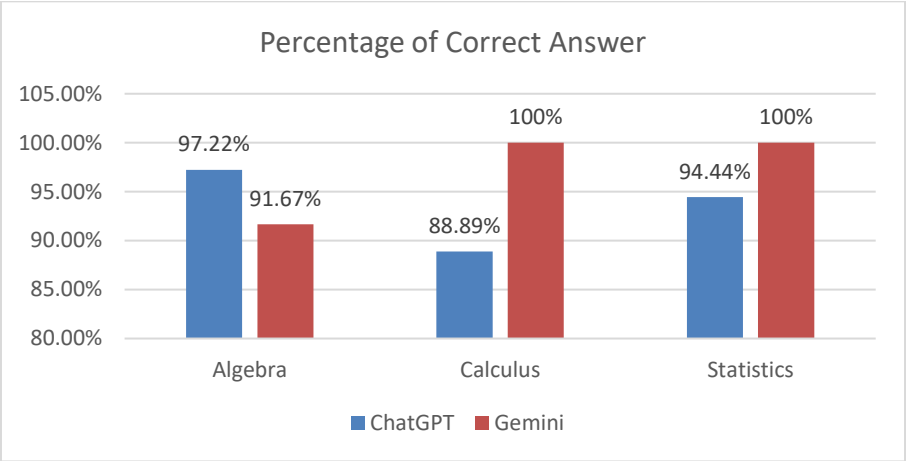

Figure 1: Comparative accuracy rates of ChatGPT and Gemini across three mathematical domains.

Figure 1 presents the percentage of correct answers obtained by ChatGPT and Gemini across three mathematical domains: Algebra, Calculus, and Statistics. Overall, both models demonstrate high levels of accuracy, exceeding 88% in all domains. However, clear domain-specific performance differences are observed.

In Algebra, ChatGPT achieved an accuracy rate of 97.22%, outperforming Gemini, which recorded 91.67%. This finding is consistent with prior evaluations indicating that GPT-based models perform strongly in structured symbolic tasks that involve rule-based transformations and equation manipulation (Brown et al., 2020; Jahin et al., 2025). Algebra problems often follow recognizable procedural patterns, and LLMs trained on large mathematical corpora tend to generalize well in such contexts. Duan et al. (2025) similarly reported that GPT-series models exhibit stable performance in symbolic and structured reasoning tasks when intermediate steps align with common transformation patterns. The higher algebra accuracy observed for ChatGPT in this study therefore aligns with existing empirical evidence suggesting relative strength in structured symbolic domains.

In contrast, Gemini demonstrated superior performance in Calculus and Statistics, achieving 100% accuracy in both domains, while ChatGPT obtained 88.89% in Calculus and 94.44% in Statistics. The substantial gap in Calculus, exceeding 11 percentage points, suggests stronger procedural consistency by Gemini in this dataset. This observation is supported by comparative benchmark studies showing that Gemini models perform competitively in computation-heavy tasks and structured reasoning benchmarks (DesignTalks, 2024; Jahin et al., 2025). Procedural calculus tasks such as differentiation and integration often require the correct application of standard formulas, and LLMs can achieve high accuracy when problem-solving pathways are well-defined (Duan et al., 2025). Gemini's

perfect accuracy in Statistics also aligns with findings that some models exhibit strong performance in numerical and formula-based computations under controlled problem conditions (Jahin et al., 2025).

Despite the high correctness rates, previous research cautions that final answer accuracy does not necessarily guarantee logically valid intermediate reasoning. Boye and Moëll (2025) demonstrated that large language models may arrive at correct mathematical answers while still containing subtle logical inconsistencies in step-by-step explanations. Similarly, Edwards (2025) highlights that simulated reasoning models can produce outputs that appear coherent yet fail to reflect genuine deductive processes. Therefore, while Gemini achieved 100% correctness in Calculus and Statistics in this study, such results should be interpreted within the broader context of known reasoning limitations in LLMs.

The domain-specific variation observed in Figure 1 supports the broader conclusion that LLM mathematical performance is context-dependent rather than uniformly robust. Lin et al. (2022) argue that language models optimize probabilistic token prediction rather than formal logical reasoning, which can lead to variability across domains. John (2025) further emphasizes that LLMs often display surface-level accuracy that deteriorates when problem structures deviate from familiar training distributions. The differences between ChatGPT and Gemini across Algebra, Calculus, and Statistics in this study reinforce the argument that model effectiveness depends heavily on task structure and problem type.

From an educational perspective, the consistently high accuracy levels may create a perception of reliability, potentially encouraging overreliance among students. Urban et al. (2024) found that students frequently accept AI-generated solutions without verifying intermediate steps, while Ateeq et al. (2024) warn that such dependence can undermine critical thinking and academic integrity. Given that both models achieved high percentages across domains, educators must emphasize validation of reasoning processes rather than sole reliance on final answers.

Based on the findings, the results demonstrate that both ChatGPT and Gemini exhibit strong mathematical performance, yet their strengths differ by domain. ChatGPT shows relatively stronger performance in Algebra, consistent with prior findings on symbolic reasoning capabilities. Gemini demonstrates superior performance in Calculus and Statistics within this dataset, aligning with benchmark studies reporting competitive performance in procedural and computation-heavy tasks. These findings support the growing body of literature indicating that LLM performance in mathematics is domain-sensitive and should be interpreted with careful attention to reasoning validity rather than final answer correctness alone.

5.0 Conclusion

This study examined the comparative accuracy of ChatGPT and Gemini in solving mathematical problems across three core domains: Algebra, Calculus, and Statistics. Using a descriptive quantitative approach, the findings show that both AI systems achieved consistently high accuracy, exceeding 88% in all tested domains. This outcome indicates that contemporary large language models can perform effectively on structured university-level mathematics tasks. The results also demonstrate clear domain-specific variation. ChatGPT performed more accurately in Algebra, suggesting relative strength in symbolic manipulation and structured transformation tasks. Gemini, in contrast, achieved perfect accuracy in Calculus and Statistics within the tested dataset, indicating strong consistency in procedural and formula-based computation. This domain-sensitive pattern aligns with prior evaluations reporting that LLM performance differs depending on task structure and the extent of multi-step procedural demands (Duan et al., 2025; Jahin et al., 2025).

However, high final-answer correctness should not be interpreted as proof of robust mathematical reasoning. The literature consistently cautions that LLMs may generate correct answers while still exhibiting hidden logical gaps, incomplete justifications, or unstable intermediate reasoning (Boye & Moëll, 2025; Edwards, 2025). Because LLM outputs are generated probabilistically, reliability may decline when problems are reworded or when the task deviates from familiar patterns (Lin et al., 2022; John, 2025). From an educational standpoint, these findings suggest that ChatGPT and Gemini can serve as useful computational assistants and learning supports, but they should not replace independent reasoning and verification. High apparent accuracy may encourage overreliance and reduce students' attention to validating solution logic, which can weaken conceptual learning and raise integrity concerns (Urban et al., 2024; Ateeq et al., 2024).

Taken together, the findings demonstrate that both ChatGPT and Gemini exhibit strong mathematical problem-solving capability, although their performance differs across domains. Future research should expand the number and difficulty of problems, include proof-oriented questions, and evaluate step-level validity to provide a deeper assessment of whether accuracy reflects genuine reasoning or surface-level pattern matching.

References

Ateeq, A., Alzoraiki, M., Milhem, M., & Ateeq, R. A. (2024). Artificial intelligence in education: Implications for academic integrity and the shift toward holistic assessment. *Frontiers in Education, 9*, Article 1470979. https://doi.org/10.3389/feduc.2024.1470979

Boye, J., & Moëll, B. (2025). *Large language models and mathematical reasoning failures*. arXiv. https://doi.org/10.48550/arXiv.2502.11574

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & others. (2020). *Language models are few-shot learners*. arXiv. https://doi.org/10.48550/arXiv.2005.14165

DesignTalks. (2024). ChatGPT-4 vs Gemini Ultra: In-depth comparison. *DesignTalks*. https://designtalks.co.za/chatgpt-4-vs-gemini-ultra-in-depth-comparison/

Duan, Y., Gong, S., & Huang, S. (2025). *Evaluation of mathematical capabilities of GPT-4, Claude, DeepSeek, LLaMA, and Gemini (DIKWP framework)*. Technical report. https://doi.org/10.13140/RG.2.2.28216.25604

Edwards, B. (2025, April 25). New study shows why simulated reasoning AI models don't yet live up to their billing. *Ars Technica*. https://arstechnica.com/ai/2025/04/new-study-shows-why-simulated-reasoning-ai-models-dont-yet-live-up-to-their-billing/

Jahin, A., Zidan, A. H., Zhang, W., Bao, Y., & Liu, T. (2025). *Evaluating mathematical reasoning across large language models: A fine-grained approach*. arXiv. https://doi.org/10.48550/arXiv.2503.10573

John. (2025). Researchers question AI's "reasoning" ability as models stumble on math problems with trivial changes. *JWSheetMetal*. https://jwsheetmetal.com/index.php/2024/04/01/researchers-question-ai-s-reasoning-ability-as/

Lin, S., Cahyawijaya, S., Lee, N., Dai, W., Su, D., & Wilie, B. (2023). *TruthfulQA: Measuring how models mimic human falsehoods*. arXiv. https://doi.org/10.48550/arXiv.2302.04023

Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education*. https://doi.org/10.1016/j.compedu.2024.105031

eISBN 978-629-98755-9-8

9 786299 875598

SIG CS@e-Learning
Unit Penerbitan
Jabatan Sains Komputer & Matematik
Universiti Teknologi MARA Cawangan Pulau Pinang

e-ISBN : 978-629-98755-x-x

Design of the cover powered by
https://www.free-powerpoint-templates-design.com/