Mathematical Capabilities of LLMs

Rahul Thakur, Christ (Deemed to be University), Delhi-NCR, Ghaziabad, India, [rahul.thakur@mca.christuniversity.in](mailto:rahul.thakur@mca.christuniversity.in)

Ramesh Chandra Poonia, Christ (Deemed to be University), Delhi-NCR, Ghaziabad, India, [rameshchandra.poonia@christuniversity.in](mailto:rameshchandra.poonia@christuniversity.in)

***Abstract*—** **Large language models (LLMs) have demonstrated potential across multiple domains, yet their performance in mathematical problem-solving remains insufficiently explored. This research evaluates five leading LLMs—ChatGPT, Gemini Pro, Qwen 110B, Mistral Large, and Mistral-8x7b—on a set of 50 mathematical problems spanning calculus, algebra, geometry, number theory, and probability. The study assesses the accuracy of their solutions and evaluates their ability to provide correct intermediate steps. A primary dataset was created to compare LLM performances, and the models were ranked based on their ability to correctly solve problems. The evaluation highlights the limitations of current LLMs in mathematical reasoning and problem-solving, while also identifying areas where improvements are needed. Future research aims to refine the dataset and continue monitoring the progression of LLM capabilities in solving increasingly complex mathematical problems.**

***Index Terms*—** **LLM benchmarking, mathematical reasoning, dataset evaluation, model comparison, LLM decision making, problem-solving**

1. Introduction

The rise of large language models (LLMs) in the field of Artificial Intelligence has opened new possibilities for Natural Language Processing, text generation, and complex problem-solving. These models, trained on extensive datasets of text and code, have demonstrated remarkable capabilities across a variety of tasks—ranging from generating human-like text to performing creative reasoning and even answering intricate questions. Recent advancements in LLMs, pioneered by models like GPT-3, have catalyzed a wave of research, pushing for continuous improvements and refinements. These developments have led to increasingly complex models capable of achieving greater performance across more diverse challenges. Serving as research platforms, cutting-edge LLMs like Gemini Pro, Mistral Large, ChatGPT, Mistral-8x7b, and Qwen 110B represent the frontier of AI technology, continuously expanding the horizons of what can be achieved in language understanding and reasoning.

Although LLMs are increasingly valued across various fields, their potential as mathematical reasoning tools has not been thoroughly examined. Mathematical problem-solving is a domain where LLMs face significant challenges due to the precise calculations and logical steps required to reach solutions. Mathematics involves a combination of formal logic, complex calculations, and structured reasoning, which is fundamentally different from tasks driven by linguistic patterns or memorization. Therefore, assessing how well LLMs can handle structured mathematical reasoning presents a critical test of their overall problem-solving ability. This research seeks to address the fundamental question: Can current large language models accurately solve mathematical problems, particularly in fields like calculus, algebra, geometry, number theory, and probability.

This paper explores the mathematical reasoning capabilities of five prominent LLMs—ChatGPT, Qwen 110B, Gemini Pro, Mistral Large, and Mistral-8x7b—through a comprehensive evaluation of 50 mathematical problems. Each model's performance was assessed based on its ability to provide correct solutions and intermediate steps across five mathematical categories. The focus is on two main issues: (i) the effectiveness of LLMs in generating correct answers to mathematical problems, and (ii) their ability to follow logical processes that lead to valid solutions. This research sheds light on the specific strengths and limitations of LLMs in tackling complex mathematical challenges.

Motivation and Significance

The drive to automate and improve mathematical reasoning through artificial intelligence is the primary motivation for this study. Mathematics, known for its rigor and structured problem-solving processes, stands to benefit significantly from AI solutions that can enhance accuracy, reduce error rates, and optimize workflows. The ability of LLMs to solve mathematical problems accurately and consistently could revolutionize several fields, including scientific research, education, and applied mathematics, where precise reasoning is crucial. However, the limitations of current models in handling logical reasoning and calculations raise important questions about their true potential in this domain

This research is significant for several reasons:

* Research gap: While many studies have explored how LLMs perform in language-based tasks like text generation and natural language processing, few have empirically examined their mathematical reasoning capabilities. This study addresses this gap by thoroughly evaluating the strengths and bottlenecks of LLMs in a mathematical setting, offering valuable insights into which types of problems they can handle and where they struggle.
* Implications for mathematics and AI: The findings presented in this research have implications for the broader mathematical and AI communities. By identifying both the strengths and weaknesses of LLMs in mathematical problem-solving, this research provides critical information that can inform the design of future models and their applications in mathematical fields.
* Contributions to AI research: This work contributes to the growing body of research on the limitations of LLMs when applied to tasks requiring structured, logical reasoning beyond language processing. The results should guide future research into the development of more specialized LLMs, better equipped to handle diverse challenges, including mathematics.

Challenges of Applying LLMs to MATHEMATICS

Applying LLMs to mathematical problems presents several unique challenges, stemming from both the nature of the models and the mathematical domain itself:

* Mathematical reasoning and calculation: LLMs, which are trained primarily on textual data, often struggle with complex mathematical calculations and reasoning. The challenges of mathematical problem-solving—requiring deep understanding of logical steps, computations, and formal proofs—demand more than the pattern-recognition abilities that LLMs excel at in language-based tasks..
* Domain-specific terminology: The field of mathematics has its own specialized language, rules, and notations that LLMs must understand to solve problems correctly. This requires models to have specific training in mathematical concepts, something that is often lacking in general-purpose LLMs.
* Explainability and interpretability: In mathematical problem-solving, it is critical to understand not just the answer but the reasoning that leads to it. LLMs, which often operate as "black boxes," lack transparency in their decision-making processes, making it difficult to verify the correctness of their solutions, especially in fields like mathematics where accountability is key.

Previous Research and Related Work

While research on LLMs in mathematical reasoning is still developing, several studies have explored the application of LLMs in related fields such as symbolic reasoning and logic. These studies provide valuable insights into the strengths and weaknesses of LLMs when applied to structured reasoning tasks. For example, previous research has shown that LLMs perform well on tasks involving symbolic manipulation, such as integration, but struggle with more complex reasoning that involves multiple steps or logical deductions. This research builds on those findings by evaluating a wider array of mathematical tasks and providing a more comprehensive picture of LLM performance in different categories of mathematics.

Furthermore, research on the limitations of LLMs in logical reasoning has identified several approaches to improve performance. These include integrating external tools like symbolic solvers, adding modules specifically designed for mathematical reasoning, and training models on specialized datasets. This study leverages these insights to design a primary dataset of mathematical problems, evaluating multiple LLMs to determine their strengths and limitations in this field.

Research Objectives and Contributions

This research aims to achieve the following objectives:

1. Performance evaluation of LLMs on mathematical problems: This research investigates the performance of five top-performing LLMs—ChatGPT, Qwen 110B, Gemini Pro, Mistral Large, and Mistral-8x7b—on a set of 50 mathematical problems across five categories. The goal is to assess how well these models generate accurate solutions and reasoning steps.
2. Understanding reasoning in LLMs: The research analyzes how these LLMs process complex problems, focusing on their logical reasoning and calculation capabilities. Identifying where these models excel or fall short helps highlight areas for improvement in model development.
3. Comparative performance analysis: By comparing the performance of these LLMs across multiple categories, this study aims to elucidate the factors that contribute to differences in their abilities to solve mathematical problems.
4. Tracking LLM progression over time: The research evaluates whether LLM performance has improved over time by using a longitudinal approach, assessing models at different stages of development to observe trends in mathematical reasoning capabilities.

The key contributions of this research are:

* + Comprehensive evaluation of LLM capabilities in mathematics: This research fills an important gap by providing a thorough assessment of how well current state-of-the-art LLMs perform across a range of mathematical tasks.  
    • Identification of strengths and weaknesses: Through a structured comparison, the study highlights specific areas where LLMs excel or face challenges, particularly in logical reasoning and computation.  
    • Insights into LLM generalizability: This research contributes to the broader understanding of how general-purpose LLMs can be applied to specialized tasks, such as mathematics, identifying the challenges and opportunities for their future development..

1. Related Work
2. Frieder et al. (2023) provide an in-depth analysis of the mathematical capabilities of ChatGPT. They evaluated ChatGPT on both publicly available datasets and hand-crafted ones specifically created to assess mathematical reasoning. The results indicated that ChatGPT often understands mathematical questions but fails to provide correct solutions, particularly in more advanced mathematical domains. This aligns with the findings of this research, where similar inconsistencies were observed when testing ChatGPT on tasks like calculus, algebra, and number theory. The work demonstrates a clear gap in the ability of current LLMs to solve mathematical problems accurately, highlighting the need for a robust evaluation framework as developed here.
3. Hendrycks et al. (2021) introduced the MATH dataset, a challenging set of problems aimed at testing LLMs' mathematical reasoning abilities. The dataset has been instrumental in benchmarking models like GPT-3 and Minerva. Despite improvements, these models only achieved moderate success rates on complex problems, typically around 50%. This reflects the difficulties that LLMs face when applied to sophisticated mathematical reasoning tasks. In the current research, it was observed that models such as Qwen 110B and Gemini Pro similarly struggled with higher-level tasks, particularly in areas like probability and calculus, reaffirming the challenges highlighted by the MATH dataset.
4. Amini et al. (2019) developed the MathQA dataset to assess LLM performance on mathematical word problems. Their findings revealed that LLMs, including those trained on large-scale data, generally underperform when faced with tasks requiring not just calculation but logical reasoning and problem comprehension. This research mirrors these conclusions, as all five LLM models evaluated—ChatGPT, Qwen 110B, Gemini Pro, Mistral-8x7b, and Mistral Large—struggled with problems involving intricate reasoning, such as those in the number theory and geometry categories. This highlights the consistent gap in LLM capabilities for real-world mathematical problem-solving.
5. Lample and Charton (2019) demonstrated the potential of transformer models in symbolic mathematics, where they outperformed traditional symbolic solvers like Mathematica and Maple. While their results show that transformers can handle specific types of mathematical problems, such as symbolic integration, the findings in this research suggest that this performance does not necessarily extend to broader categories of mathematics. Even transformer-based models like Mistral Large and Mistral-8x7b showed significant shortcomings in basic algebra and calculus tasks, indicating that while LLMs excel in symbolic tasks, their general mathematical reasoning abilities remain limited.
6. Cobbe et al. (2021) tested LLMs on the GSM8K dataset, which focuses on elementary-level math problems. The study found that models like GPT-3 struggled even with relatively simple arithmetic tasks, with success rates falling below 60%. This finding is consistent with the observations in this research, where even the larger, more advanced models such as Qwen 110B and Gemini Pro often failed in providing correct answers to fundamental problems in algebra and probability. These results highlight the current limitations of LLMs across both elementary and advanced mathematical domains, supporting the conclusion that significant improvements are necessary for more accurate mathematical reasoning.
7. Lewkowycz et al. (2022) examined the ability of LLMs to perform quantitative reasoning, focusing on models like GPT-3 and its successors. Their work revealed that, while LLMs can often follow the structure of mathematical problems, they frequently make computational errors or logical leaps. This is directly relevant to this study, as similar issues were observed across all five models when tested on problems in calculus and geometry. These logical and computational gaps reinforce the need for specialized datasets, like the one created here, to measure the specific weaknesses of LLMs in mathematical problem-solving.
8. Wiegreffe et al. (2023) evaluated InstructGPT’s ability to generate coherent mathematical proofs. They found that, although the model could generate valid proof structures, it often made critical logical errors that undermined the accuracy of the final solution. This mirrors the findings in the geometry and number theory categories of this research, where models like Gemini Pro and Mistral-8x7b failed to provide accurate proofs despite understanding the overall problem structure. The dataset developed in this study further highlights this issue by offering a structured comparison of multiple LLMs, demonstrating that logical consistency in mathematical reasoning is a persistent challenge for current models.
9. Piékos et al. (2021) explored how models like BERT perform when faced with mathematical reasoning tasks, finding that the model often fails to maintain logical consistency across steps. This finding aligns with this research, as models like ChatGPT and Mistral-8x7b demonstrated similar inconsistencies when solving algebraic and calculus problems. The dataset used here emphasizes these logical gaps, providing a quantitative measure of how frequently these models provide incomplete or incorrect answers to structured mathematical questions. These results suggest that while LLMs can process mathematical language, they struggle with the underlying logic required to solve complex problems.
10. Davies et al. (2021) highlighted the potential of AI-assisted intuition in advancing mathematical research. Their work demonstrates that AI can aid in discovering new mathematical insights but still falls short in executing precise, logical problem-solving in standard mathematical tasks. This relates closely to the findings here, where LLMs like Qwen 110B and Gemini Pro often provided creative but ultimately incorrect solutions to algebra and geometry problems. The need for accurate logical reasoning in LLMs remains a critical challenge that this study aims to address through the creation of a primary dataset designed specifically to test these limitations.
11. Le Scao et al. (2022) introduced BLOOM, a multilingual language model designed for broad applicability across tasks. However, their work did not specifically address mathematical problem-solving, a critical gap that this research seeks to fill. In the evaluation of LLMs, it was found that even large-scale models like Qwen 110B struggle with mathematical tasks, particularly in areas like probability and calculus. By focusing on the mathematical capabilities of LLMs, this research builds on the foundational work of BLOOM by providing targeted insights into how these models perform in specialized domains like mathematics.
12. Lobo (2023) commented extensively on the public excitement surrounding ChatGPT's capabilities while also pointing out its consistent shortcomings in mathematical reasoning. This aligns with the observations made during the evaluation of ChatGPT and other LLMs, where despite its general proficiency in natural language tasks, ChatGPT often fails to provide accurate answers to more structured, technical mathematical problems. This highlights the persistent gap between public perception and actual model performance in mathematical domains, a gap that this research directly addresses through comprehensive testing and analysis.
13. Rabe et al. (2020) focused on the potential of LLMs to engage in formalized mathematical reasoning through the use of large databases of formal proofs. While their work demonstrates progress in formalized domains, the findings here show that current LLMs like Gemini Pro and Mistral-8x7b still struggle with natural language-based mathematical problem-solving. This distinction between formal and informal mathematical reasoning highlights a significant gap in the capabilities of LLMs that this research aims to quantify and address through targeted evaluations in areas like algebra and number theory.
14. Kung et al. (2022) explored the use of LLMs in medical education, demonstrating that while LLMs perform well on standardized tests, they often struggle with tasks requiring deeper logical reasoning. This finding resonates with the results of this research, where models like ChatGPT and Qwen 110B provided reasonable answers to standard, formulaic mathematical problems but failed when more complex reasoning was required. The dataset created for this research helps to quantify these limitations, providing a structured approach to evaluating how LLMs perform when faced with tasks requiring both computational accuracy and logical consistency.
15. Harrison et al. (2014) examined the history and progress of interactive theorem proving and its implications for AI models. While theorem-proving systems have advanced significantly, LLMs like ChatGPT and Mistral-8x7b are still far from achieving similar levels of performance in natural language-based mathematical tasks. This research directly addresses this gap by challenging LLMs with a series of structured problems across various domains of mathematics. The results demonstrate that while LLMs can sometimes emulate basic theorem proving, they often fail to produce the detailed logical steps necessary for more complex proofs.
16. Lample and Charton (2019) demonstrated that deep learning models can outperform traditional symbolic mathematics solvers in specific tasks, such as symbolic integration. However, the findings in this research suggest that this performance does not extend to all areas of mathematics. Even models like Mistral-8x7b and Gemini Pro, which are based on transformer architectures, struggled with tasks in algebra and geometry. These results indicate that while LLMs have potential in symbolic tasks, their broader mathematical reasoning capabilities still require significant improvement, a gap that this research highlights through targeted evaluation.
17. Methodology

This section outlines the methodology used in this research to evaluate the performance of selected Large Language Models (LLMs) in solving mathematical problems. The methodology includes the selection of LLMs, construction of the mathematical problem set, data collection process, and evaluation metrics for assessing model performance.

1. *Selection of LLMs*

This study involves the selection of five state-of-the-art LLMs (Fig. 1), chosen for their architectural diversity and prowess in various language understanding and generation tasks. These models, further detailed in Section III, represent a spectrum of capabilities, allowing for a comprehensive evaluation across the mathematical domain.

ChatGPT: Developed by OpenAI, this model is widely known for its robust conversational abilities and natural language understanding.

Gemini Pro: A multimodal model from Google AI, known for its high performance in complex language understanding and reasoning tasks.

Mistral Large: Renowned for its efficiency in code generation and reasoning, this model comes from a leading AI startup in the field of language modeling.

Mistral-8x7b: A more advanced variant from the Mistral family, featuring enhanced capabilities in problem-solving and efficient use of parameters.

Qwen-110b: A highly capable model developed by Alibaba, known for its large-scale language model architecture designed for advanced reasoning and complex tasks.

1. *Construction of the Mathematical Problem Set*

The problem set used in this research consists of 50 complex mathematical problems, covering a range of topics including algebra, calculus, and number theory. The problems were carefully selected from [insert source], ensuring that they required a high level of reasoning and the application of various mathematical formulas.

Each problem was clearly defined, with detailed requirements, numerical data, and a known solution provided. This allowed for a standardized evaluation across all models. To maintain consistency, data cleaning was performed by removing extraneous spaces and ensuring a uniform format for all problems and their solutions.

1. *Data Collection Process*

The mathematical problems were posed to each LLM, and their responses were recorded systematically. The questions were administered using the [insert platform] to ensure uniformity in the interaction with each model. The following parameters were used to guide the model responses:

Temperature: 0.7 (to introduce a degree of variability and creativity in responses).

Top p: 1 (to consider the full spectrum of potential token options).

Max Tokens: 2000 (to allow for complete, well-explained responses).

A multi-turn prompting strategy was utilized to ensure that each LLM fully understood the problems. This approach involved an initial question, followed by clarifying prompts to guide the models toward more precise solutions

Manual Evaluation

Data Analysis

Limitations

Evaluation Metrics

Construction of Mathematics Problem Set

Data Collection Process

Selection of LLMs

Start

End

1. *Evaluation Metrics*

The performance of the models was evaluated using three key metrics:

Steps/Formula Accuracy: This metric assesses how accurately each LLM applies the steps and formulas needed to solve the mathematical problems. A score of 1 is given for correct application, and 0 for incorrect methods.

Numerical Output: This metric measures the correctness of the final numerical answer generated by the model. If the answer is accurate, the model receives a score of 1; otherwise, a score of 0 is assigned.

Overall Score: The overall score is calculated by summing the scores for Steps/Formula Accuracy and Numerical Output, giving a comprehensive measure of the model’s performance..

1. *Manual Analysis*

To ensure the accuracy of the evaluation, all model outputs were manually reviewed and compared against the correct solutions. This manual process involved verifying each step taken by the models, the formulas they used, and their final outputs to determine their correctness. Although manual evaluation introduces a degree of subjectivity, it remains the most thorough approach for assessing the models’ problem-solving capabilities..

1. *Data Analysis*

The collected data was analyzed using Python, with various statistical tools used to generate graphs that compare the performance of the LLMs across different problem categories. Additionally, a regression analysis was conducted to identify performance trends and predict how each model might improve over time. Visualizations were generated to illustrate the comparative strengths and weaknesses of each model in solving mathematical problems.

Fig. 1: Flowchart of Research Methodology

1. *Limitations*

Several limitations should be noted in this research:

Limited Mathematical Scope: The study focuses on a specific set of mathematical problems, which may not fully represent the models' capabilities across a broader range of tasks.

Subjectivity in Manual Evaluation: While manual evaluation allows for thorough assessment, it introduces potential biases and human error.

Limited Number of LLMs: The evaluation covers five LLMs, which, while representing diverse architectures, may not fully capture the performance of other models available on the market.

Despite these limitations, this research provides valuable insights into the problem-solving capabilities of LLMs, highlighting their strengths and areas for improvement in handling complex mathematical tasks.

.

Results and Discussion

In this section, we demonstrate the performance of our selected LLMs on accounting problem set in two separation times. We examine the same 4 statistics on overall scores and steps for each model including mean score, standard deviation of scores, minimum score, maximum value. What it does is assess the quality and level of reasoning every LLM uses, making that possible only by discussing there reasons as one.

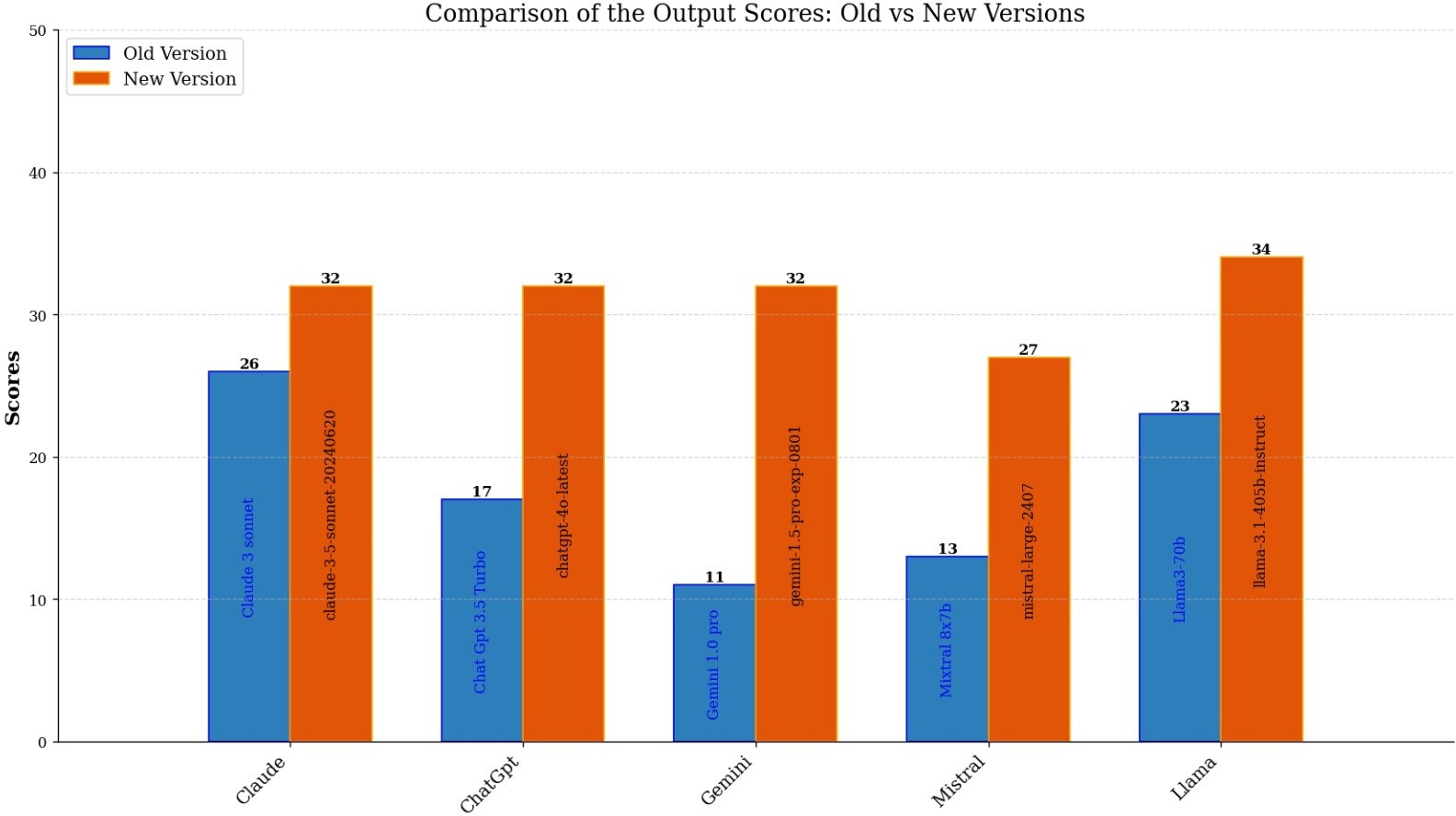
*Insights for Analysis*

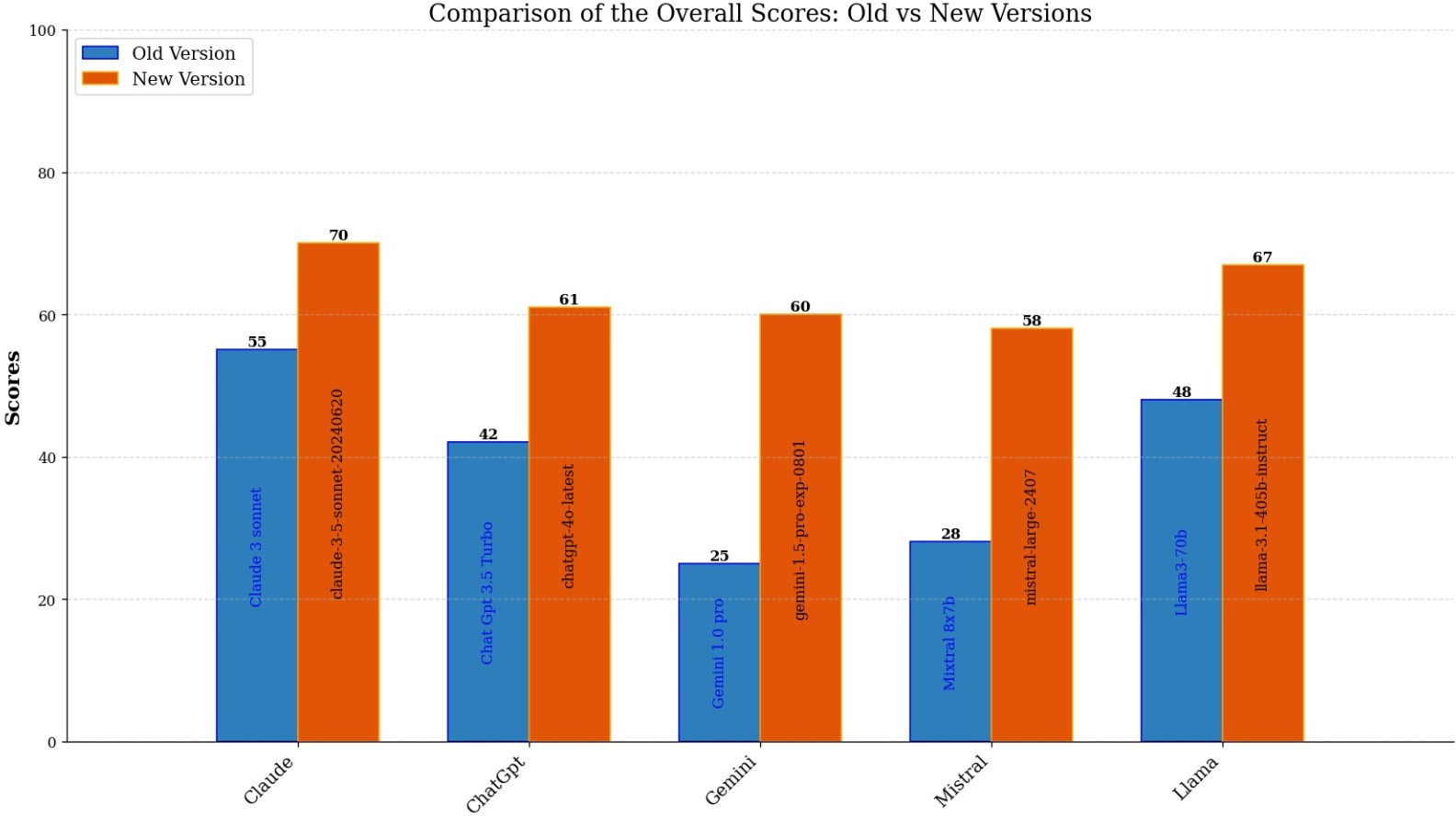
Our analysis is centred on a couple of key factors:

* Model Accuracy and Consistency: In our analysis, we compare both time periods mean scores versus standard deviations to determine the most accurate/reliable models between them.
* Handling Complex Queries: We analyze the performance of each model on challenging queries, focusing on their ability to provide accurate solutions and detailed reason- ing.
* Comparison with Ideal Responses: We compare the model outputs with ideal responses to assess the quality and completeness of their solutions.
* Performance on Different Question Categories: We ex- amine the performance of each model on different types of accounting problems, such as computational versus conceptual questions, to identify potential strengths and weaknesses.

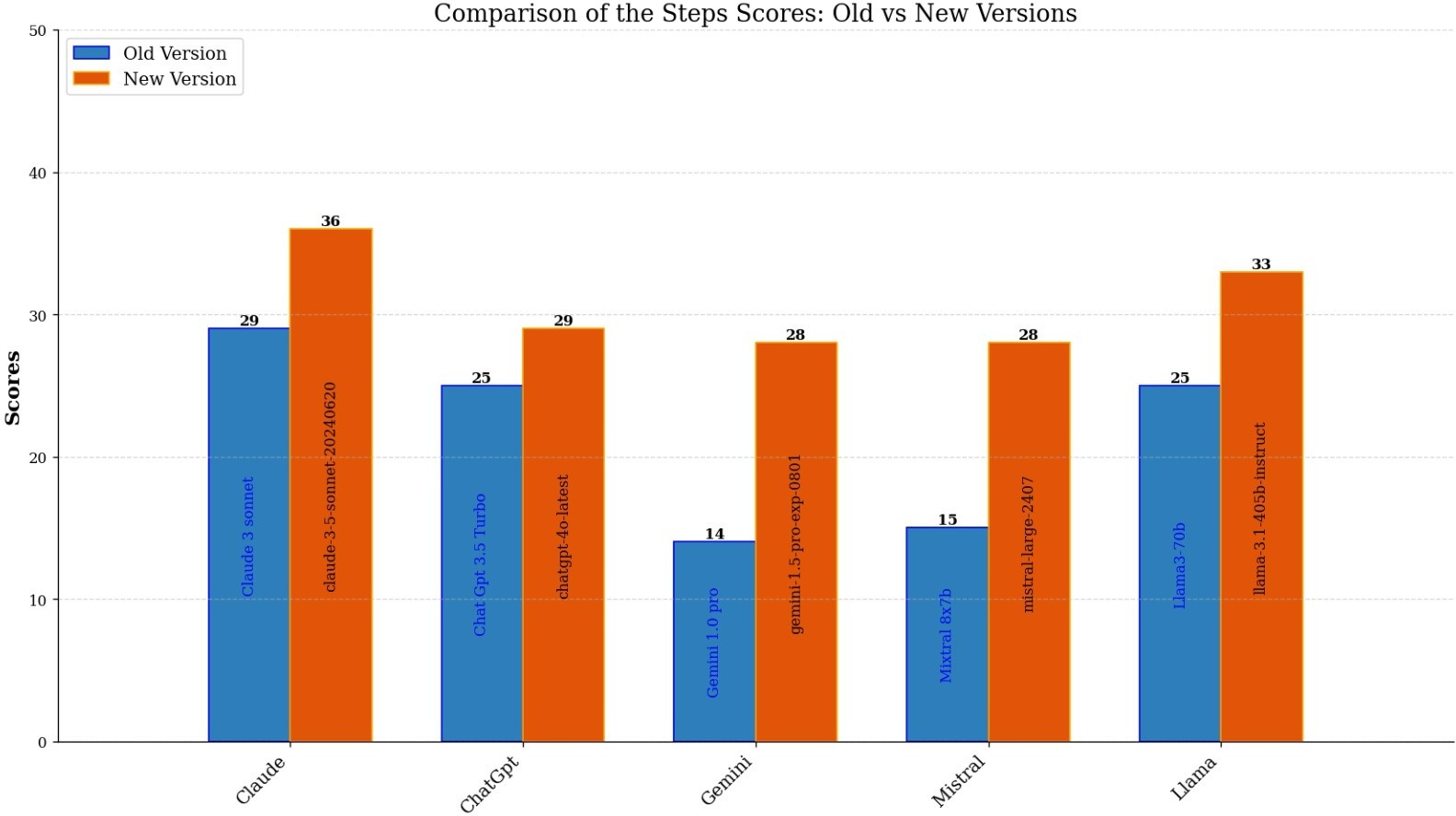
*Summary of Comparative Performance*

We assess each LLM’s performance using the following metrics:

* **Mean**: The average score or quantity of steps the model takes on all questions combined. The mean is calculated as:



* 1. Overall Scores



* 1. Steps Scores

Σ*N xi*

Mean = *i*=1

*N*

where *N* is the total number of questions (50 in this case), and *xi* is the score or number of steps for question *i*.

* **Standard Deviation (Std)**: This calculates the variability or dispersion of the scores or steps from the mean. It shows how dispersed the values are. The standard deviation is given by:

(c) Output Scores

Fig. 2: Scores of Models

TABLE I: Summary of performance metrics for the older versions of the LLMs.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **Claude**  **3** | **GPT 3.5** | **Gemini**  **1.0** | **Mixtral**  **8x7b** | **Llama 3** |
| Mean Score | 1.10 | 0.84 | 0.50 | 0.56 | 0.96 |
| Std Score | 0.89 | 0.89 | 0.81 | 0.86 | 0.92 |
| Min Score | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Max Score | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 |
| Mean Steps | 0.58 | 0.50 | 0.28 | 0.30 | 0.50 |
| Std Steps | 0.50 | 0.51 | 0.45 | 0.46 | 0.51 |
| Min Steps | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Max Steps | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

s Σ*N*

*i*=1 *i*

*N −* 1

Standard Deviation =

(*x −* Mean)2

* Minimum : the smallest score of the model. same as time, the least number of sourced texts that the model needs to click to select the correct one.
* Max: the largest score the model could get as well. The most steps taken by the model for all the questions

*Analysis*

*Score Comparison:* All models in the new dataset exhibit improved mean scores, indicating enhanced accuracy in their newer versions. The standard deviation for scores remains relatively consistent, suggesting that the models’ consistency has generally improved or remained stable.

*Steps Comparison:* The mean number of steps has increased for most models in the new dataset, suggesting a more thorough approach to problem-solving. The standard deviation of steps remains similar, indicating a consistent increase in the depth of processing across questions.

TABLE II: Summary of performance metrics for the newer versions of the LLMs.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **Gemini**  **1.5** | **Claude**  **3.5** | **Mistral**  **Large** | **GPT 4o** | **Llama**  **3.1** |
| Mean Score | 1.20 | 1.40 | 1.16 | 1.22 | 1.34 |
| Std Score | 0.88 | 0.83 | 0.91 | 0.89 | 0.82 |
| Min Score | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Max Score | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 |
| Mean Steps | 0.56 | 0.72 | 0.56 | 0.58 | 0.66 |
| Std Steps | 0.50 | 0.45 | 0.50 | 0.50 | 0.48 |
| Min Steps | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Max Steps | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

*Growth Analysis: Changes in Scores and Steps Between Old and New Models*

We calculated the percentage growth in mean score and mean steps for each model, comparing their performance in the old and new datasets. The results are summarized below:

TABLE III: Summary of percentage growth in scores and steps for each model between the old and new datasets.

|  |  |  |
| --- | --- | --- |
| **Model Comparison** | **Score Growth** | **Steps Growth** |
| Claude 3 vs. Claude 3.5 | 27.27 | 24.14 |
| Chat GPT 3.5 vs. Chat GPT 4o | 45.24 | 16 |
| Gemini 1.0 vs. Gemini 1.5 | 140 | 100 |
| Mixtral 8x7b vs. Mistral Large | 107.14 | 86.67 |
| Llama 3 vs. Llama 3.1 | 39.58 | 32 |

*Summary of Growth*

*Score Growth:* All models saw a noteworthy increase in their average scores with Gemini 1.5 leading the growth at 140% and Mistral Large coming second, growing by 107.14%. Each of these represents significant improvements in terms of the accuracy and finally, model performance!

*Steps Growth:* The general growth of the average steps taken by different models is evidence to, an increasingly planned and careful manner of solving problems. Gemini 1.5 has highest steps growth at 100% that corresponds to its significant score improvement as well.

1. Conclusion

This research is significant in that it clearly demonstrates tha tLLMs are getting to grips with increasingly sophisticated accounting matters. Finally, all performance and reasoning accomplishments of models have evolved between those two time periods in a statistically significant way—so LLMs are developing very rapidly on financial accounting tasks. Nonetheless, the results of this study suggest that advances are encouraging despite continuing problems in terms for processing complex calculations and elaborate accounting principle explanations. The result of the research is a strong reinforcement that LLMs could soon be essential for accountants, assuming all routine tasks requiring high levels of accuracy and financial data insights. It goes without saying that, in years to come and as new models get built-out developed refined finalised etc., we can a priori expect they should only become more powerful strong tools for handling advanced financial accounting. By answering this

question empirically, the study paves a detailed path for further research on LLMs in accounting and thus marks an important contribution to our collective understanding of artificial intelligence.

References

1. G. Team, “Gemini: A family of highly capable multimodal models,” 2024.
2. S. Karamcheti, L. Orr, J. Bolton, T. Zhang, K. Goel, A. Narayan,

R. Bommasani, D. Narayanan, T. Hashimoto, D. Jurafsky, C. D. Manning,

C. Potts, C. Re´, and P. Liang, “Mistral - a journey towards reproducible language model training,” 2021.

1. A. Dubey, “The llama 3 herd of models,” 2024.
2. Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi,

C. Wang, Y. Wang, *et al.*, “A survey on evaluation of large language models,” *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

1. T. B. Brown, “Language models are few-shot learners,” 2020.
2. Y. Ge, W. Hua, K. Mei, J. Tan, S. Xu, Z. Li, Y. Zhang, *et al.*, “Openagi: When llm meets domain experts,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
3. Menghao, L. Minggao, Y. Fengchao, L. Yagang, and ShiYajing, “Research on the application of llm in power finance middle platform,” in *2023 2nd Asian Conference on Frontiers of Power and Energy (ACFPE)*, pp. 282– 289, 2023.
4. P. Srivastava, M. Malik, V. Gupta, T. Ganu, and D. Roth, “Evaluating llms’ mathematical reasoning in financial document question answering,” in *Findings of the Association for Computational Linguistics ACL 2024*,

pp. 3853–3878, 2024.

1. Y. Huang, X. Liu, Y. Gong, Z. Gou, Y. Shen, N. Duan, and W. Chen, “Key-point-driven data synthesis with its enhancement on mathematical reasoning,” *arXiv preprint arXiv:2403.02333*, 2024.