Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.22871

Research on the Problem-Solving Ability of Large Language Models in Mathematics Questions

Zehan Zhou

Mechanics and Electronics, Beijing Jiaotong University, Beijing, China
23222029@bjtu.edu.cn

Abstract: In recent years, an increasing number of large language models have emerged and been extensively applied in mathematical problem - solving. Their remarkable computational capabilities not only enable the rapid calculation of complex equations but also provide in - depth logical analysis, offering substantial assistance to humans dealing with intricate mathematical challenges. However, the vast diversity of mathematical domains, ranging from algebraic number theory to differential geometry, demands highly specialized training frameworks and unique datasets tailored to specific problem types. This poses multifaceted challenges, such as data sparsity in niche areas and the need for domain - specific knowledge integration. This study will primarily investigate three key areas: a comprehensive overview of large language models, including their underlying architectures; their application status in various mathematical domains, like how they are used in financial mathematics for risk assessment; and the factors influencing the quality of mathematical reasoning approaches, such as the impact of training data size. Additionally, this research will summarize existing superior training methods, such as fine - tuning techniques, and provide directions for the future development of large language models in mathematical problem - solving, for example, exploring more efficient neural network architectures.

Keywords: Large Language Models, mathematical problem solving, transformer, training data

1. Introduction

How to solve mathematical problems has always been a topic that humans in every era aspire to explore. With the advancement of technology, each era has new tools to assist humanity in advancing mathematics. In recent years, the rapid development of large language models has also provided new approaches for mathematical reasoning. Large language models have been proven to have immense capabilities in solving complex problems [1], and their computational and information retrieval abilities have played a significant role in mathematical reasoning. However, large language models are not omnipotent in mathematical reasoning [2]. Different fields have different demands on large language models, and no single model or training method can make large language models adaptable to all demands. This paper, based on existing research, will analyze the factors affecting the accuracy of large language models and the capabilities and deficiencies required for large language models to solve different types of mathematical problems, such as arithmetic, geometry, and math application problems.

DOI: 10.54254/2755-2721/2025.22871

2. Overview of large language models

Large language models are artificial intelligence models developed to understand, generate, and process human language. By being trained on large - scale text data, these models learn the relationships and logic within text, thereby enabling them to communicate with humans. The current core model of large language models is the transformer [3], which employs an encoder-decoder architecture. The encoder maps the input sequence to a fixed-length representation, while the decoder generates the output sequence. The primary innovation of the transformer over traditional models lies in the use of self-attention mechanisms, which eliminate the need for traditional convolutional networks. This design enables the model to focus on different positions of the input sequence simultaneously, rather than relying on sequential time steps, thus achieving more efficient parallel processing.

Pre-training and fine-tuning are critical processes in deep learning for improving model efficiency and generalization ability. During the pre-training phase, models learn fundamental features, such as language syntax and image texture, through unsupervised or self-supervised learning on large-scale general datasets (e.g., text and images). This phase establishes a general parameter foundation for subsequent tasks. Fine-tuning, based on pre-trained models, involves adjusting parameters using small-scale labeled data from specific domains. It adapts to downstream tasks through full parameter updates, partial parameter freezing, or efficient technical adjustments, significantly reducing data and computational requirements. In addition to pre - training and fine - tuning, the process of training models also encompasses steps such as data preprocessing, model architecture design, and hyperparameter optimization. These processes improve model accuracy and specificity and serve as important complements to pre - training and fine - tuning.

3. Current situation of application of large language model in mathematical problem solving

Large Language Models (LLMs) have been extensively applied in diverse aspects of mathematical computations and reasoning, such as solving algebraic equations, proving geometric theorems, handling mathematical word problems, and other related mathematical tasks.

3.1. Arithmetic

In the domain of algebraic equation solving, Large Language Models (LLMs) demonstrate marked proficiency when addressing elementary problems such as those in the GSM8K[4]. However, persistent limitations emerge when confronting mathematical competition-level problems, necessitating substantial human guidance to achieve satisfactory solutions. This highlights two critical observations:

1. Computational Robustness: LLMs exhibit formidable computational capabilities in executing algorithmic procedures and formulaic manipulations.

2. Contextual Reasoning Deficits: Their capacity to autonomously parse complex problem structures, identify latent constraints, and synthesize multi-step reasoning pathways remains deficient. Specifically, challenges persist in extracting critical information from linguistically intricate problem statements and formulating adaptive solution strategies when confronted with non-standard mathematical scenarios.

This performance gap underscores the current boundaries of LLMs in sophisticated reasoning tasks, suggesting that while their pattern recognition and procedural execution are highly developed, their abstract reasoning mechanisms require further architectural innovations.

DOI: 10.54254/2755-2721/2025.22871

3.2. Geometric theorem proving

Geometric proof constitutes a critical domain in mathematics, requiring sophisticated logical reasoning and the application of domain-specific symbolic systems. While Large Language Models have demonstrated partial competence in proving fundamental theorems, significant challenges persist in achieving comprehensive understanding and autonomous generation of proofs [5].

The first challenge is complexity of Geometric Symbolic Systems, the complexity of geometric symbolic systems presents notable challenges. Geometric proofs depend on specialized symbols and terminology that are highly specific to the domain. The variations in symbolic representations found across different textbooks and research literature can complicate both the training and interpretation of models. Therefore, to effectively adapt large language models, dedicated training is needed to enable them to recognize and accurately employ geometric symbols.

Geometric reasoning also necessitates visual inputs, which offer indispensable visual information (such as shapes, lengths, angles) to directly facilitate problem understanding and the formulation of proof strategies. Visual inputs can substantially enhance LLMs' accuracy in geometric theorem proving by enabling spatial and relational comprehension.

The generation of complex proofs presents significant challenges. High - level geometric proofs frequently demand abstract thinking, the creative generation of lemmas, and the capacity to synthesize logic across multiple steps—capabilities that are still underdeveloped in current large language models. Additionally, performance deteriorates when models are confronted with non-routine problems that demand innovative proof strategies, as these tasks go beyond the rote application of known theorems.

The application of LLMs to geometric theorem proving holds considerable potential but remains constrained by symbolic ambiguity, visual dependency, and limitations in abstract reasoning.

3.3. Math word problems

Mathematical word problems, which depict scenarios through textual or verbal descriptions instead of explicit equations, demand a dual proficiency in natural language understanding and mathematical reasoning. While Large Language Models have made strides in tackling these problems, their error rates remain significantly higher compared to tasks involving direct equation manipulation [6]. This discrepancy underscores the unique complexities of integrating linguistic interpretation with algorithmic problem-solving.

A structured analysis of the challenges reveals several key issues that large language models face in mathematical reasoning. Firstly, ambiguity within linguistic contexts presents a significant challenge since word problems frequently contain ambiguous phrasings, implicit assumptions, or culturally - specific references. For instance, a problem like "a train leaving Station A at 60 mph" assumes knowledge of units and linear motion, which may not be universally understood. Additionally, LLMs may misinterpret contextual cues or fail to resolve underspecified variables, further complicating accurate problem interpretation.

Furthermore, multi-step reasoning with latent constraints often proves difficult for models. Solutions frequently require synthesizing multiple mathematical concepts, such as combining ratios, algebra, and geometry in a single problem. However, LLMs struggle to identify intermediate steps or infer unstated constraints, such as assuming integer solutions or non-negative quantities, which are often critical to reaching the correct solution.

Another challenge arises from the symbolic translation of text to equations. Converting narrative descriptions into formal mathematical expressions demands precise semantic parsing, as errors often arise from misalignments between linguistic constructs and their mathematical counterparts. This misalignment can lead to incorrect equations and, consequently, erroneous solutions.

DOI: 10.54254/2755-2721/2025.22871

Finally, domain - specific knowledge gaps pose a significant obstacle. Problems in specialized domains, like physics or finance, demand familiarity with jargon, conventions, and domain - specific reasoning. The lack of such knowledge in LLMs can hinder their ability to accurately interpret and solve problems within these contexts. Together, these challenges highlight the need for improved guidance, data diversity, and advanced mechanisms to enhance the mathematical reasoning capabilities of LLMs.

4. Factors that affect the quality of ideas for solving large language models

The quality of problem-solving approaches in large language models is influenced by a multitude of factors, which can be broadly categorized into internal and external dimensions. Internal factors include the number of model layers and attention mechanisms, while external factors encompass training data and methodologies.

4.1. Internal dimension

In terms of model depth, deep models demonstrate substantial advantages over shallow models. Deep models, which are characterized by architectures featuring multiple hidden layers, have the capacity to learn more intricate information and features. Moreover, deep models necessitate less human intervention since they can independently extract features from raw data, thus endowing them with the ability to adapt to a broader spectrum of scenarios and applications. However, shallow models, with their low computational costs and data requirements, remain advantageous in specific contexts.

Attention mechanisms refer to the process of selectively focusing on particular aspects of information to enhance efficiency. Certain attention mechanisms, such as sparse attention, may prioritize efficiency by neglecting less critical information. However, the primary challenge in attention mechanisms lies in their difficulty in identifying long-range dependencies between words. To address this limitation, self-attention mechanisms have been developed. These mechanisms compute the relevance of each piece of information relative to others, assigning different weights to different elements. This approach facilitates a deeper understanding of key information within problems and enables the generation of more rational problem-solving steps.

4.2. External dimension

The influence of training data on models is obvious. Abundant high - quality data can remarkably enhance accuracy, and diverse datasets enable large language models to adapt to different types of problems.

The way users interact with the system significantly influences the accuracy of the responses. In general, providing clear guidance and incorporating key terms in inquiries can substantially enhance correctness. For example, the Chain - of - Thought (CoT) method [7] accomplishes this by presenting multiple examples, allowing the model to gradually generate a reasoning chain that structures the solution process and promotes the derivation of conclusions. Similarly, the REAP framework (Reflection, Explicit Problem Deconstruction, and Advanced Prompts) [8] employs a reflective approach to guide the decomposition of queries into manageable components and generates relevant contextual information to strengthen the problem-solving process. Different prompting strategies yield varying effects and are applicable to domain-specific problem-solving tasks.

As mentioned before, large language models sometimes lack access to common knowledge that is well - known to humans. This constraint can cause the model to provide incorrect answers in a way that seems coherent and semantically consistent. Enhancing the quality of the provided context represents one potential solution [9], as improving the relevance between contextual elements allows the model to extract information more effectively. Additionally, manually providing supplementary

DOI: 10.54254/2755-2721/2025.22871

information serves as another method to increase accuracy. Furthermore, establishing a feedback mechanism enables the gradual correction of the large language model's vulnerabilities.

5. Conclusion

This study on the mathematical problem-solving capabilities of large language models analyzed both their contributions and limitations in various mathematical domains. Based on this research, when utilizing large language models to address relatively abstract and complex problems, more human guidance and prompts are necessary, along with the employment of more diverse datasets. Additionally, the implementation of advanced attention mechanisms and the provision of higher-quality contextual information could enhance the model's ability to extract relevant information effectively. It is hoped that this research will contribute to the future development and training of large language models, fostering the advancement of more comprehensive mathematical reasoning capabilities across various fields. Ultimately, the objective is to endow large language models with diverse problem - solving capabilities in mathematics, instead of relying on specific training frameworks and guiding methods, thus better aiding humans in mathematical problem - solving.

References

[1] *Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog,M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2023. Mathe matical discoveries from program search with large language models. Nature, pages 1–3*

[2] *Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, Wenpeng Yin (2024) Large Language Models for Mathematical Reasoning: Progresses and Challenges. https://arxiv.org/pdf/2402.00157*

[3] *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Att ention is all you need. Advances in neural information processing systems, 30.*

[4] *Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plap pert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, John Schulman. (2021) Training Ve rifiers to Solve Math Word Problems. https://arxiv.org/pdf/2110.14168*

[5] *Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He & Thang Luong. (2024) Solving olympiad geometry without human demonstrations. Nature,625:476–482.*

[6] *Mingyu Zong and Bhaskar Krishnamachari. (2023). Solving math word problems concerning systems of equations with GPT-3. In Proceedings of AAAI, pages 15972–15979.*

[7] *Kathrin Seßler, Yao Rong, Emek Gözlüklü, Enkelejda Kasneci.(2024) Benchmarking Large Language Models for Math Reasoning Tasks. https://arxiv.org/pdf/2408.10839*

[8] *Ryan Lingo, Martin Arroyo, Rajeev Chhajer.(2024).Enhancing LLM Problem Solving with REAP: Reflection, Explicit Problem Deconstruction, and Advanced Prompting. https://arxiv.org/pdf/2409.09415*

[9] *Xiang Li, Haoran Tang, Siyu Chen, Ziwei Wang, Anurag Maravi, Marcin Abram.(2023) Context Matters: Data-Efficient Augmentation of Large Language Models for Scientific Applications. https://arxiv.org/pdf/2312.07069*