A Review of the Integration of Basic Large Language Models and Mathematical Reasoning Capabilities: Taking the Application of GPT Series in Mathematical Theorem Proving and Equation Solving as an Example

Ruihao Cheng
Jinan Xinhang Experimental Foreign Language School, Jinan, Shandong, China, 250000
c_rh1@icloud.com

Abstract

This article focuses on the integration research of basic large language models and mathematical reasoning capabilities, taking the application of the GPT series models in mathematical theorem proving and equation solving as the entry point. By sorting out the development trajectory of the GPT series models, analyzing the key technological breakthroughs in their application in the field of mathematics, exploring their specific performance in mathematical theorem proving and equation solving, revealing the current challenges faced by fusion research, and looking forward to the future development direction, the aim is to provide theoretical references for promoting the further development of large language models in the field of mathematical reasoning.

Keywords: Basic large language model; Mathematical reasoning ability; GPT series; Proof of mathematical theorems; Equation solving.

1. Introduction

Mathematics, as the cornerstone of human cognition, supports many fields such as natural science, engineering, medicine, finance, computer science and social science [1]. Since the early 1960s, developing computational models capable of independently solving mathematical application problems has been an important research direction in the field of natural language processing (NLP) [2]. This pursuit is not only about solving arithmetic and algebraic expressions, but also aims to promote the development of general reasoning mechanisms, which are key to achieving artificial general intelligence (AGI) [3]. By overcoming the complexity of mathematical reasoning, researchers strive to expand the logical deductive ability of AI systems and their understanding and manipulation of symbolic knowledge.

In recent years, the rise of pre-trained language models (PLMs) and large language models (LLMs) has driven breakthroughs in this field [4]. Models such as BERT, RoBERTa [5], BART [6], GPT-1 [7], and GPT-2 have demonstrated outstanding language and numerical reasoning capabilities by learning from large-scale text corpora. Especially LLMs represented by the GPT series, with their huge parameter numbers and powerful context learning capabilities, have made remarkable progress in complex tasks such as proving mathematical theorems and solving equations [8]. For instance, the performance of GPT-4 in mathematics competition problems has approached the level of human experts, while GPT-5 Pro has independently derived mathematical conclusions that are more accurate than the original text and provided a complete proof process [9].

However, although LLMs have shown great potential in mathematical reasoning, there is still a fundamental contradiction between their core mechanism (generating the most likely sequence based on probability distribution) and the deterministic, logical and abstract requirements of mathematical reasoning [10]. Therefore, in-depth research on the integration mechanism of LLMs and mathematical reasoning capabilities not only helps to enhance the performance of models in the field of mathematics, but also provides an important reference for promoting the development of AI systems towards a more reliable and interpretable direction.

2. The Development Trajectory of the GPT Series Models

2.1 Laying the Foundation for the Early GPT Model

The Generative Pre-trained Transformer (GPT) series of models originated from the Transformer architecture, which effectively captures long-distance dependencies in text through the self-attention mechanism, providing a foundation for the model to handle complex language tasks. Early GPT models, such as GPT-1 and GPT-2, mainly learned the general representation of language through large-scale unsupervised pre-training and demonstrated certain capabilities in natural language generation and understanding tasks. These models laid the technical and data foundation for the subsequent development of the GPT series, allowing people to see the potential of large language models in handling various language tasks.

2.2 The Scale and Capability Leap of GPT-3

The emergence of GPT-3 is an important milestone in the development of the GPT series. The number of its parameters has significantly increased to 175 billion. Through training with massive amounts of data, the model has demonstrated strong zero-shot and few-shot learning capabilities. In terms of mathematical reasoning, GPT-3 has begun to be capable of handling some simple mathematical problems, such as basic arithmetic operations and simple logical reasoning. However, its performance still has significant limitations and is not yet competent for complex mathematical theorem proving and equation solving. However, the scale effect and powerful language generation ability of GPT-3 provide new directions and ideas for subsequent research.

2.3 Deepening and Expansion of GPT-4 and Subsequent Versions

GPT-4 further optimizes the model architecture and training methods on the basis of GPT-3, enhancing the performance and stability of the model. It has significantly improved its mathematical reasoning ability, is capable of handling more complex mathematical problems, and has achieved outstanding results in some mathematical benchmark tests. Since then, the GPT series has been continuously iterated and updated, such as versions like GPT-5, achieving all-round technological breakthroughs in architectural innovation, reasoning capabilities, security mechanisms, and API flexibility. GPT-5 adopts an embedded trinity integrated architecture, consisting of an efficient response model for handling daily queries, a deep reasoning model for solving complex tasks, and an intelligent routing system serving as the "brain" of the architecture for real-time decision-making. It can automatically select the optimal model for processing based on task types, significantly enhancing the model's performance in complex tasks such as mathematical reasoning.

3. Key Technological Breakthroughs in the Application of GPT Series Models in Mathematical Reasoning

3.1 Chain-of-Thought Prompting, CoT

The thought chain prompt is one of the important techniques for enhancing the mathematical reasoning ability of the GPT series models. The core idea is to add examples of intermediate reasoning steps in the prompts, guiding the model to output the step-by-step thinking process before generating the final answer. This technology breaks down complex reasoning tasks into a series of seemingly simpler local next-token prediction tasks, making the model behave as if it were "thinking". For instance, when solving mathematical problems, by providing detailed solution steps for similar problems as prompts, the model can imitate these steps for reasoning, thereby significantly improving performance on benchmarks such as GSM8K. However, the steps for generating CoT are essentially still based on the most common text sequences in the training data that are related to the problem pattern. The rigor of its logic is not intrinsically guaranteed, and sometimes it may produce seemingly reasonable but actually incorrect "illusory" reasoning chains.

3.2 Instruction Fine-tuning, SFT

Instruction fine-tuning fine-tunes the pre-trained model by using high-quality "instruction-response" data, making the model more adept at "playing" the role of a mathematical problem solver and learning standard solutions to specific types of problems. For instance, works such as MetaMath, WizardMath, and MathInstruct have significantly enhanced the ability of models to solve mathematical problems following instructions by fine-tuning on a large amount of mathematical instruction data. Instruction fine-tuning enables the model to better understand the mathematical problems raised by users and generate solutions in accordance with specific formats and requirements, thereby enhancing the accuracy and reliability of the model in mathematical reasoning tasks.

3.3 Reinforcement Learning with Process Supervision, RL with PRM

Reinforcement learning and process supervision are key technologies for further enhancing the reliability of mathematical reasoning in the GPT series of models. Unlike traditional result supervision (which only focuses on whether the final answer is correct), process supervision pays attention to the correctness of each step in the problem-solving process. Human annotators will conduct fine-grained scoring of the reasoning steps. The model optimizes itself by learning human preferences for answer quality (correctness, logic, clarity) and using reinforcement learning algorithms (such as PPO) Generate outputs that are more likely to yield higher rewards, that is, logically coherent and correct reasoning paths. For instance, works such as Math-Shepherd and OmegaPRM have reduced the high annotation costs through automated process supervision, expanded the application scope of process supervision, and effectively alleviated the "seemingly reasonable" erroneous reasoning that models may generate.

3.4 Tool Use

Due to the inherent flaws of the GPT series models in precise calculations (such as large number multiplication and floating-point operations) and complex symbol derivations (such as solving higher-order equations and integrals), tool-using techniques have emerged. This technology enables the model to learn to invoke external tools (such as calculators, Python interpreters, WolframAlpha, etc.) to handle tasks that these models are not good at. The model performs computations or queries by generating API calls or code snippets in a specific format, and then integrates the returned results back into its natural language inference stream. For instance, frameworks such as Toolformer, ART, TORA, and ReTool have demonstrated the immense potential of tool usage, liberating the GPT series of models from the pressure of being "all-knowing and all-powerful", allowing them to focus on their areas of expertise in semantic understanding and task decomposition, and outsourcing precise calculations and symbolic operations to reliable professional tools. This effectively enhances the model's ability to solve complex mathematical problems.

4. Application of GPT Series Models in Mathematical Theorem Proving

4.1 Breakthrough of the New Theorem Independently Proved by GPT-5 Pro

In 2025, OpenAI researchers revealed a discovery that shocked the academic community: After reading a mathematical paper, GPT-5 Pro independently derived a more accurate mathematical conclusion than the original text and provided a complete proof process. This paper studies an important issue in the field of convex optimization - whether the resulting optimization curve has convexity when using the gradient descent algorithm to optimize smooth convex functions. GPT-5 Pro provides a more accurate threshold for one of the boundary problems than the original text and offers a rigorous mathematical proof process. This discovery means that AI is no longer merely a reteller or organizer of human knowledge but has truly acquired the ability to think independently and reason innovatively, marking an important turning point in the development of AI.

4.2 GPT-5 Pro Demonstrates the Uniqueness of the Approach

The proof approach of GPT-5 Pro is completely different from that of anthropologists. In the proof of the above-mentioned convex optimization problem, it ingeniously employs two fundamental inequalities of convex L-smooth functions: the Bregman divergence inequality and the standard cocoercive inequality. Through exquisite algebraic operations, it successfully further refines the convexity conditions, demonstrating profound mathematical knowledge and innovative thinking. Anthropologists, on the other hand, utilized the Bregman divergence inequality of convex L-smooth functions to establish inequalities for three different point pairs respectively. Then, they summed these inequalities with different weights and ingeniously simplified the complex gradient terms through identities, ultimately obtaining precise mathematical boundaries. This difference proves that AI is not merely a simple imitation or plagiarism but truly possesses the ability to independently explore and innovate.

4.3 Limitations of the GPT Series in Mathematical Theorem Proving

Although the GPT series models have made certain breakthroughs in the proof of mathematical theorems, there are still many limitations. On the one hand, when the model is confronted with complex theorem proofs, it may still encounter problems such as illogical rigor and reasoning errors. For instance, incorrect formulas or reasoning steps might be used during the proof process, leading to a wrong final conclusion. On the other hand, the model is highly dependent on the training data. When encountering problems that exceed the training distribution, its performance will drop sharply. Furthermore, the innovation level of the current GPT series models in mathematical theorem proving still needs to be

improved. Although they can provide some new conclusions and proofs, there is still a considerable gap compared with the major breakthroughs of human mathematicians.

5. Application of GPT Series Models in Equation Solving

5.1 Performance of Early GPT Models in Solving Simple Equations

Early GPT models have already demonstrated certain capabilities when dealing with simple equation solving problems. For instance, for simple linear and quadratic equations with one variable, the GPT model can usually correctly find the solutions. Users only need to clearly and explicitly pose questions and use the correct mathematical terms and symbols, and the GPT model can generate solutions to the equations based on the knowledge it has pre-trained. However, when it comes to higher-level algebra, such as systems of linear equations with multiple variables or polynomial equations, early GPT models may not be able to solve them correctly. In addition, when dealing with problems that require symbolic operations or the use of special solution methods, the GPT model may also encounter difficulties.

5.2 Advantages of GPT-5 in Solving Complex Equations

With the continuous development of the GPT series models, GPT-5 has shown significant advantages in solving complex equations. In the 2025 AIME test in the United States, GPT-5 scored 94.6% without tools and 99.6% with the collaboration of Python tools. For the most challenging and complex tasks, GPT-5 Pro Professional Edition scored 100% after using Python. This is attributed to the powerful language comprehension ability, reasoning ability and tool usage technology of GPT-5. It can accurately understand the complex equation problems raised by users, and through internal reasoning and the invocation of external tools, such as the Python interpreter, perform precise calculations, thereby arriving at the correct solutions.

5.3 Challenges and Responses in Equation Solving Applications

In the application of equation solving, the GPT series models also face some challenges. For instance, for some equations containing special symbols or complex structures, the model may not be able to correctly recognize and parse them. In addition, the numerical accuracy issues that may be involved in the equation solving process can also affect the accuracy of the model's solution. To address these challenges, researchers have been constantly optimizing the training data and algorithms of the model to enhance its ability to handle special symbols and complex structures. At the same time, by integrating tool usage techniques and using professional mathematical software for precise calculations, the accuracy of equation solving can be ensured.

6. Challenges Currently Faced by Fusion Research

6.1 The Contradiction between the Intrinsic Mechanism of the Model and the Demand for Mathematical Reasoning

The core mechanism of basic large language models is to generate the most likely sequences based on probability distributions, and mathematical conclusions are usually deterministic, with extremely high requirements for logic and abstraction. This contradiction in the internal mechanism makes the model face difficulties when dealing with complex mathematical problems. For instance, the model might generate reasoning steps that seem reasonable but are actually incorrect, or make jumps during the logical reasoning process, resulting in an inaccurate final conclusion.

6.2 Issues Regarding Data Dependency and Generalization Capabilities

The performance of the GPT series models is highly dependent on the quality and quantity of the training data. Although large-scale data training can enable models to learn rich mathematical knowledge and problem-solving patterns, when encountering problems that exceed the training distribution, the performance of the model will drop sharply. In addition, the model may have overfitting in the training data, resulting in poor performance on new test data and insufficient generalization ability.

6.3 A balance between Logical Rigor and Innovation

In mathematical reasoning, logical rigor is of vital importance. However, the current GPT series models find it difficult to guarantee the rigor of logic when generating reasoning processes. Meanwhile, although the model can provide some new conclusions and proofs, its degree of innovation still needs to be improved. How to enhance the innovativeness of the model on the basis of ensuring logical rigor is an important challenge currently faced by fusion research.

7. Prospects for Future Development Directions

In the future, researchers will continue to optimize the architecture and algorithms of the GPT series models to better meet the demands of mathematical reasoning. For instance, explore new attention mechanisms to enhance the model's ability to capture long-distance dependencies; Design more effective training objective functions to enhance the logical reasoning ability of the model; By integrating technologies such as knowledge graphs, more abundant mathematical knowledge and structured information are provided for the model to enhance its performance.

With the development of multimodal technology, integrating various modal information such as text, images, and audio into the GPT series models is expected to further enhance the mathematical reasoning ability of the models. For instance, when solving geometric problems, integrating image information can help the model better understand the geometric structure and spatial relationship of the problem. In addition, the application of the GPT series models in the field of mathematical reasoning will also expand to more cross-disciplinary scenarios, such as physics, chemistry, engineering, and other fields, providing stronger support for solving practical problems.

In the future, human-machine collaboration will become an important model for promoting the development of mathematical reasoning. Human mathematicians and the GPT series of models can complement each other and leverage their respective strengths. Human mathematicians can provide professional mathematical knowledge and innovative thinking, guiding models to conduct more in-depth reasoning and exploration. The GPT series of models can quickly handle large amounts of data and complex calculations, providing valuable references and inspirations for human mathematicians. Through human-machine collaboration, it is expected to achieve collaborative innovation and promote major breakthroughs in the field of mathematics.

8. Conclusion

The integration of basic large language models and mathematical reasoning capabilities is one of the current research hotspots in the field of artificial intelligence. As a typical representative of basic large language models, the GPT series models have achieved certain results in areas such as proving mathematical theorems and solving equations. Through breakthroughs in key technologies such as thought chain prompts, instruction fine-tuning, reinforcement learning and process supervision, and tool usage, the mathematical reasoning ability of the models has been continuously enhanced. However, current fusion research still faces challenges such as the contradiction between the intrinsic mechanism of the model and the demand for mathematical reasoning, the issue of data dependence and generalization ability, and the balance between logical rigor and innovation. In the future, it is necessary to continuously optimize the model architecture and algorithms, promote multi-modal integration and cross-domain applications, enhance human-machine collaboration and collaborative innovation, so as to facilitate the deep integration of basic large language models and mathematical reasoning capabilities, and make greater contributions to the development of artificial intelligence and the progress of the mathematical field.

REFERENCES

[1] Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., ... & Che, W. (2025). Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. arXiv preprint arXiv:2503.09567.

[2] Hosseini, M. J., Hajishirzi, H., Etzioni, O., & Kushman, N. (2014, October). Learning to solve arithmetic word problems with verb categorization. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 523-533).

[3] Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. Communications of the ACM, 64(7), 58-65.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[6] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[7] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
[8] Zong, M., & Krishnamachari, B. (2023). Solving math word problems concerning systems of equations with GPT models. Machine learning with applications, 14, 100506.
[9] Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023). Mathematical capabilities of chatgpt. Advances in neural information processing systems, 36, 27699-27744.
[10] Wang, P. Y., Liu, T. S., Wang, C., Wang, Y. D., Yan, S., Jia, C. X., ... & Yu, Y. (2025). A Survey on Large Language Models for Mathematical Reasoning. arXiv preprint arXiv:2506.08446.