UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Do Large Language Models Truly Grasp Mathematics? An Empirical Exploration from Cognitive Psychology

Permalink

https://escholarship.org/uc/item/24x9t7s1

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Ma, Shuoyoucheng
Xie, Wei
Wang, Zhenhua
et al.

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Do Large Language Models Truly Grasp Mathematics? An Empirical Exploration from Cognitive Psychology

Shuoyoucheng Ma[1,†], Wei Xie[1,†,✉], Zhenhua Wang[1], Xiaobing Sun[2], Kai Chen[3]
Enze Wang[1], Wei Liu[1], Hanying Tong[1]

[1]College of Computer Science and Technology, National University of Defense Technology
[2]Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR)
[3]Institute of Information Engineering, Chinese Academy of Sciences
✉Corresponding author: xiewei@nudt.edu.cn
[†]Equal contributors

Abstract

The cognitive mechanism by which Large Language Models (LLMs) solve mathematical problems remains a widely debated and unresolved issue. Currently, there is little interpretable experimental evidence that connects LLMs' problem-solving with human cognitive psychology. To determine whether LLMs possess human-like mathematical reasoning, we modified the problems used in the human Cognitive Reflection Test (CRT). Our results show that even with the use of Chain-of-Thought (CoT) prompts, mainstream LLMs, including the o1 model (noted for its reasoning capabilities), have a high error rate when solving these modified CRT problems. Specifically, the average accuracy rate dropped by up to 50% compared to the original problems. Further analysis of LLMs' incorrect answers suggests that they primarily rely on pattern matching from their training data, which aligns more with human intuition (System 1 thinking) rather than with human-like reasoning (System 2 thinking). This finding challenges the belief that LLMs have genuine mathematical reasoning abilities comparable to humans. As a result, this work may adjust overly optimistic views on LLMs' progress toward Artificial General Intelligence. Our dataset and experimental data can be accessed at https://osf.io/74yj2/.

Keywords: Larger Language Model, Chain-of-Thought, Cognitive Reflection Test

Introduction

Large Language Models (LLMs) are considered the dawn of Artificial General Intelligence (AGI) (Bubeck et al., 2023). Models such as ChatGPT (OpenAI, 2023), GPT-4 (Bubeck et al., 2023), Claude (Anthropic, 2023), Gemini (Gemini Team et al., 2024), GLM (Du et al., 2022), and o1-preview (OpenAI, 2024) have garnered considerable attention from both academia and industry. These models exhibit significant potential across various fields, including education (Pérez-Núñez, 2023), healthcare (Goyal et al., 2024; Li et al., 2024), coding (Yu et al., 2024), and social governance (Azim, 2024). This is partly attributed to the 'emergence phenomenon' (Schaeffer et al., 2023), which allows LLMs, due to their large training datasets and numerous parameters, to perform tasks they were not specifically trained for.

Using mathematical skills as an example, most LLMs have demonstrated remarkable abilities to tackle these problems. Using the Chain-of-Thought (CoT) method, the capacity of LLMs to solve mathematical problems can be further increased (Bi et al., 2024; Brown et al., 2020; Wang et al., 2023; Wei et al., 2023; Yao et al., 2023). However, due to the interpretability challenges posed by large-scale neural networks, there is still no scientific consensus on the origins and mechanisms underlying the mathematical capabilities of LLMs.

Conducting psychological measurement experiments on LLMs can help enhance the interpretability of research into LLMs' thinking. Previous research (Hagendorff et al., 2023) has demonstrated through experiments that the CoT method can effectively help LLMs handle the pitfalls in Cognitive Reflection Test (CRT) problems (Frederick, 2005). CRT problems are some well-crafted math or logic problems that human testers often get wrong(Hagendorff et al., 2023) due to intuitive thinking (System 1(Kahneman, 2011; Sloman, 1996; Stanovich, 1999; Tversky and Kahneman, 1974)). By using the CoT method, LLMs are prompted to rely more on human-like logical reasoning (System 2(Kahneman, 2011; Sloman, 1996; Stanovich, 1999; Tversky and Kahneman, 1974)), enhancing their accuracy in solving these problems. Advanced models like GPT-4 have even achieved higher accuracy than humans in these CRT tasks. However, the authors also raised speculative questions in their work, suggesting, "It is possible that some models encountered enough examples in their training to solve them 'from memory'".

We repeated and improved the experiment conducted by the previous study. If LLMs genuinely possess the intrinsic capability to comprehend mathematical logic, as is widely hypothesized, their accuracy in responding to the modified problems should not experience a marked decrement. However, our experiments revealed a distinctly opposing result: Even when employing the CoT approach, prominent LLMs, including the latest o1 model, continued to manifest a considerable error rate for the modified problems. Further analysis of the incorrect answers indicates that LLMs may not have developed a logical reasoning proficiency akin to System 2 or acquired comprehensive mathematical cognitive abilities. They predominantly resort to a methodology reminiscent of System 1, which involves matching and producing responses based on the similarity between user inquiries and training data during the text generation process. This investigation may serve to temper the overly optimistic anticipations regarding the effectiveness of CoT and the competencies of LLMs in approximating AGI.

Experiment Design

Method for Problem Modification. A single researcher manually performed all modifications to the dataset in our

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

study. To ensure the accuracy of these modifications, we also invited an additional researcher to verify the problems.

Method for Experimental Implementation. To uncover the problem-solving strategies of LLMs, we appended explicit instructions ('Please think step by step.') to the end of each problem, encouraging the use of the CoT method.

Methods for Result Analysis and Statistics. In this study, we comprehensively evaluated five prominent open-source and commercial LLMs. To ensure the accuracy of the assessment results, two independent researchers performed a detailed analysis and documentation of each model's responses to all queries. A response was considered accurate only when the problem-solving approach and the final answer were correct. After the statistical examination by the two independent researchers, a third researcher reviewed the discrepancies in their analytical results and made a final determination to ensure consistency in the results. Furthermore, to enhance the reliability of the experimental findings, three replicate trials were conducted on each dataset, with average values subsequently calculated. This methodological framework carried out all experimental procedures described in this study.

Method for Human Comparative Experimentation. We selected five researchers not involved in this study in our laboratory for this Experimentation. Each individual responded to 20% of the problems from each dataset, constituting a comprehensive examination of all datasets. We required them to provide the problem-solving process (activating system 2). The scoring methodology was consistent with that employed for LLMs. We calculated the accuracy based on these five examination papers, which served as a control group for comparison with the LLMs' results.

Experiment I: Changing the Numbers in A Problem Without Altering Its Description and Principle

We first replicated the experiment on the CRT3 data set from the study by Hagendorff(Hagendorff et al., 2023), which comprises 50 mathematical problems. We introduced three types of modifications to test problems, as delineated in Table 1. Type A represents the original problems in the CRT3 test, characterized as typical exponential growth problems, and incorporates three numbers in the problem statement. Type B modifies two of these numbers, while type C modifies all three. Type D replaces two numbers with letters, transforming the arithmetic problem into an algebraic one. To prevent LLMs from being unsure how to handle algebraic symbols, we appended a suffix for guidance: "X and Y are both numbers, you can use them to represent the final answer." Modifications of Types B, C, and D do not alter the description of the original problem (Type A). Thus, the underlying mathematical principle of the original problem remains unchanged.

The results are shown in Fig.1. In the case of Type A problems, the highest accuracy reached 100%, the lowest was 54.0%, and the average accuracy was 86.8%. For Type B problems, which involved modifications to two numbers, the accuracy significantly decreased, with the highest being 92.0%, the lowest 25.3%, and the average 68.5%($\delta = 18.3\%; \chi^2(1) = 23.00; P < 0.001$). For the problems of Type C, where three numbers were modified, the accuracy further declined, with the highest at 80.0%, the lowest at 10.7%, and the average 53.1%($\delta = 15.4\%; \chi^2(1) = 11.91; P < 0.001$). The accuracy again dropped sharply when the arithmetic problems were transformed into algebraic ones (Type D). The best-performing model, GPT-4, achieved an accuracy of only 29.3%, whereas ChatGPT 3.5 scarcely provided correct answers to any problems. The average accuracy was reduced to 20.9%($\delta = 32.2\%; \chi^2(1) = 54.00; P < 0.001$). However, the accuracy of human participants did not show a significant decrease (Fig. 1).

Table 1: Problem Types and Modifications

Type	Modification	Example
A	Original Problem	In a city, a virus is spreading, causing the total number of infected individuals to **double each day**. If it takes **6 days** for the entire city's population to be infected, how many days would it require for **half of** the people to become infected? Please think step by step.
B	Two Numbers Modified	In a city, a virus is spreading, causing the total number of infected individuals to **triple each day**. If it takes 6 days for the entire city's population to be infected, how many days would it require for **1/27 of** the people to become infected? Please think step by step.
C	All Three Numbers Modified	In a city, a virus is spreading, causing the total number of infected individuals to **triple every three days**. If it takes **36 days** for the entire city's population to be infected, how many days would it require for **1/27** of the people to become infected? Please think step by step.
D	Algebraic Transformation	In a city, a virus is spreading, causing the total number of infected individuals to **triple every 3X days**. If it takes **6Y days** for the entire city's population to be infected, how many days would it require for 1/27 of the people to become infected? X and Y are both numbers, and you can use them to represent the final answer. Please think step by step.

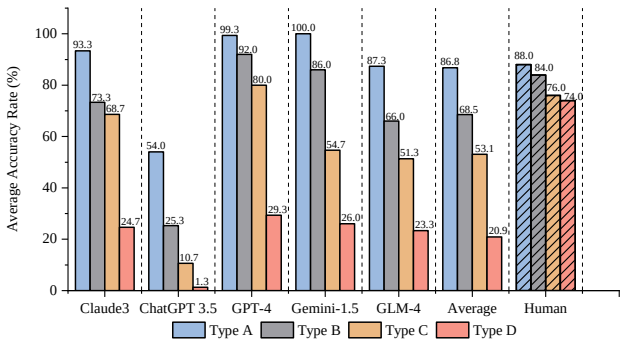


Figure 1: Accuracy of five mainstream LLMs in answering four types of problems. Human experimental results are indicated with slashes.

As shown in Fig. 2, we analyzed all incorrect answers provided by the five LLMs for each type of problem and found that in type B, errors arising from incorrect solution steps (e.g., omitting a step or altering the original calculation method) constituted 93.2%. In contrast, those only due to mathematical calculation (e.g., 16/2=4) accounted for 6.8%. In type C, errors due to incorrect solution steps accounted for 94.9%, and errors solely due to mathematical calculation accounted for 5.1%. In type D, errors due to incorrect solution steps accounted for 97.8%, and errors caused solely by mathematical calculation accounted for 2.2%. In contrast, the majority (82.1%) of human errors are found in mathematical calculation (Fig. 2). Post-analysis discussions with test-takers revealed that these errors primarily stem from the extensive number of problems, causing calculative exhaustion.

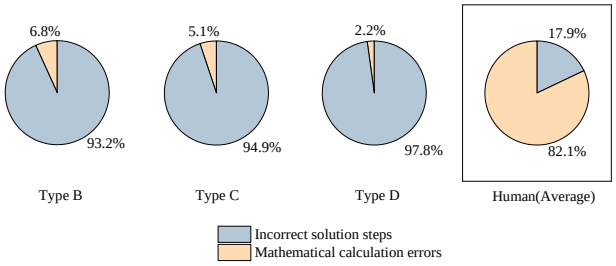


Figure 2: The proportion of incorrect answer types in answering three types of problems. Human experimental results are indicated with box.

Result Analysis of Experiment I: The modifications in types B, C, and D only altered the specific numbers in the original problems without changing the problems' mathematical principles and computational rules. For a human with mathematical and logical thinking who can solve problems by formulating equations or programming, merely changing the input numbers without modifying the mathematical principles should not significantly decrease accuracy (as shown in Fig. 1). However, the performance of LLMs differs significantly from that.

Experiment II: Modifying the Principle of A Problem While Maintaining Similarity in Its Description

We designed a reverse experiment to further substantiate the inferences drawn from Experiment I. In this experiment, we substantially altered the fundamental principles of the mathematical problems, endeavoring to preserve descriptions that closely mirrored the original versions. Subsequently, we investigated whether LLMs persisted in utilizing their problem-solving strategies on the original problem or adapted their methods to the modified problem throughout the problem-solving process. We conducted three distinct subexperiments corresponding to the three types of datasets.

For CRT1: Transforming Additively Separable Problems into Non-Additively Separable Problems

For instance, Fig. 3 illustrates the crucial distinction between an original CRT1 problem and its modified version. In the original problem, the total cost is simply the sum of the prices of two items. However, in the modified version, reaching the two items from the starting point through a shared segment must be deducted from the overall distance calculation. We use the string similarity calculation function within the difflib module of Python to quantify the degree of similarity in textual expression between the modified problem and the original, resulting in an average similarity of 75.91%. Despite the similar wording in the problem statements, the underlying mathematical principles are fundamentally distinct.

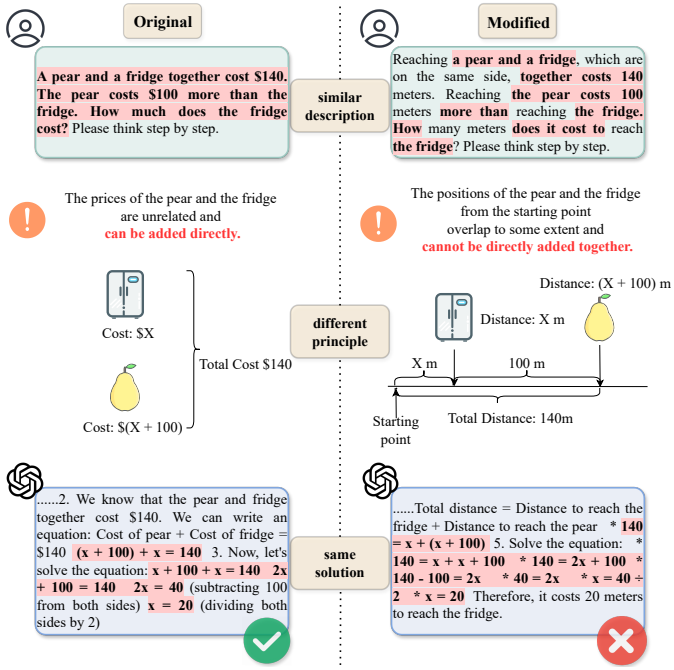


Figure 3: An example of an original CRT1 problem and its modified problem.

Fig. 4A illustrates our findings. For the original CRT1

problems, four LLMs achieved a perfect accuracy of 100%, with GLM-4 closely following at 95.3%, yielding an overall accuracy of 99.1%. However, when confronted with the modified problems, the accuracy significantly declined. GLM-4 recorded the highest accuracy rate, albeit a mere 5.3%. Claude3 and Gemini-1.5 failed to provide correct answers to any of the modified problems. On average, the five LLMs attained only 1.5% accuracy($\delta = 97.6\%; \chi^2(1) = 472.41; P < 0.001$). Human accuracy has decreased by only 14%. This means that problems modified to include overlapping components have become more complex and require more thought.

By analyzing incorrect answers, as shown in Fig. 4B, we found that 95.9% of errors were attributed to the application of the original problem's solution method to the modified problem (in the problem-solving process, add the distance directly.), whereas 4.1% resulted from new solution steps but with errors. Relatively speaking, human experimenters less frequently adopt the original problem's solution method.

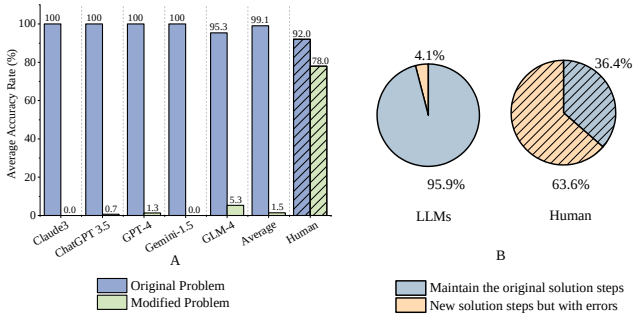


Figure 4: Accuracy of the LLMs and Human(with slash) when answering the CRT1 problems (A), and the proportion of incorrect answer types in modified CRT1 problems (B).

For CRT2: Transforming Problems Related to the Number of Individuals into Problems Independent of Individual Count

As illustrated in Fig. 5, we take the first problem in the original CRT2 dataset as an example. In the original problem, all workers work together, and the total amount of product is related to the number of workers. In the modified problem, working together is changed to walking together. Therefore, the total distance walked is independent of the number of people. We calculated the average similarity using the same function as in the previous sub-experiment, yielding 82.33%.

As illustrated in Fig.6A, the experimental findings demonstrate a significant decrease in the accuracy of various models. Claude3's accuracy plummeted from an initial level of 62.0% to 22.0% ($\delta = 40.0\%; \chi^2(1) = 14.82; P < 0.001$). Similarly, GPT-4 experienced a dramatic reduction from 94.0% to 33.3%($\delta = 60.7\%; \chi^2(1) = 37.20; P < 0.001$). Gemini-1.5 experienced the most substantial decline, dropping from 94.7% to 19.3%($\delta = 75.4\%; \chi^2(1) = 54.85; P < 0.001$). GLM-4 also encountered a considerable decrease, with its

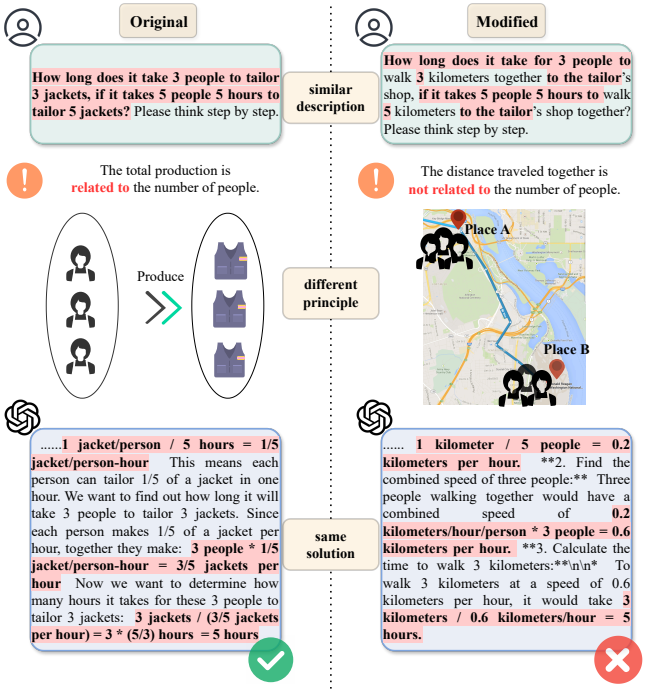


Figure 5: An example of an original CRT2 problem and its modified problem.

accuracy falling from 65.3% to 28.0%($\delta = 37.33\%; \chi^2(1) = 12.54; P < 0.001$). ChatGPT 3.5 declining from 50.7% to 34.7%($\delta = 16.0\%; \chi^2(1) = 2.00; P = 0.157$). Human experiments showed that the problems became easier.

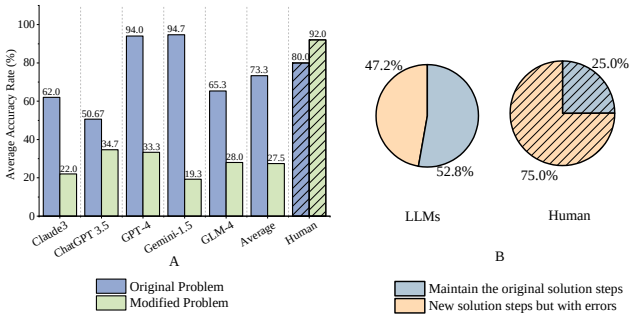


Figure 6: Accuracy of the LLMs and Human(with slash) when answering the CRT2 problems (A), and the proportion of incorrect answer types in modified CRT2 problems (B).

Upon scrutinizing all erroneous answers to the modified problems, as shown in Fig. 6B, it was observed that 52.8% of errors employed the original CRT2 methodology to address the modified problems (incorporating the number of participants into the calculation process). This indicates that, due to the similarity in wording between the two types of problems, LLMs sometimes overlook the differences in their mathematical principles and consequently choose the same solution

steps as those for the original problems, which are incorrect.

For CRT3: Transforming Exponential Growth Problems into Linear Growth Problems

Fig. 7 presents the first problem in the CRT3 dataset as an illustrative example. In the original problem, the number of viruses doubles each day, resulting in exponential growth of the total number. In contrast, the modified problem features a constant daily increase in the number of viruses, which leads to linear growth. We calculated the average similarity using the same function, yielding a result of 88.89%.

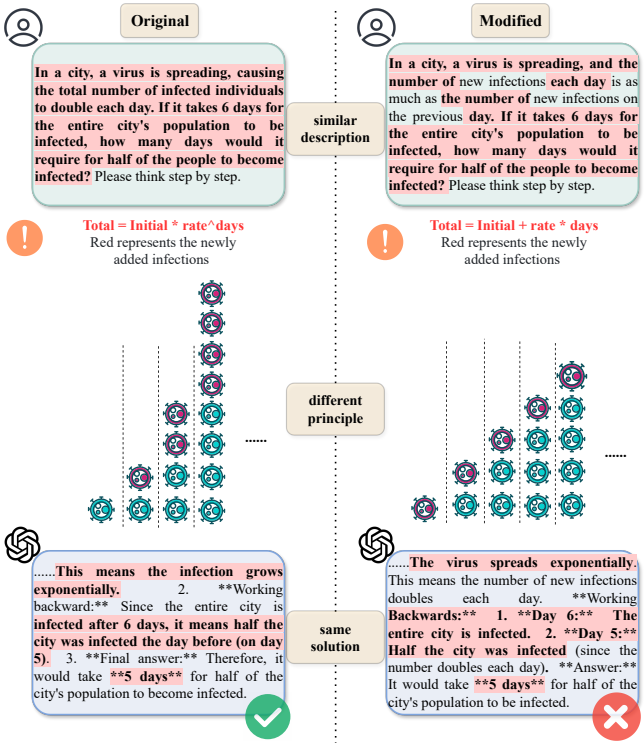


Figure 7: An example of an original CRT3 problem and its modified version.

The experimental results in Fig. 8A demonstrate that Gemini-1.5 achieved the highest accuracy of 100% for the original CRT3 problems. In contrast, ChatGPT 3.5 recorded the lowest accuracy at 54.0%, with an average accuracy of 86.8%. Following the modification of the problems, the average accuracy decreased significantly to 12.5% ($\delta = 64.3\%; \chi^2(1) = 272.83; P < 0.001$). The accuracy rate of Gemini-1.5 decreased to 20.0%($\delta = 80.0\%; \chi^2(1) = 63.38; P < 0.001$), while Claude3's accuracy was merely 1.3% ($\delta = 92.0\%; \chi^2(1) = 81.23; P < 0.001$). Conversely, humans demonstrated enhanced accuracy attributable to the reduced complexity of transforming the initial exponential problems into linear formats.

The main reason for the decline in the average accuracy is that these models continued to employ the problem-solving approach of the original problem when confronted with the modified problem. The proportions of Claude3, ChatGPT 3.5, GPT-4, Gemini-1.5, and GLM-4 that identified the modified problem as one of exponential growth were 81.7%, 77.8%, 85.7%, 72.5%, and 71.5%, respectively, although it should have been identified as a linear growth problem. Among these, the proportion of problems that were solved entirely using the original method—yielding the same answer as the original problem—constituted 37.8%, 31.8%, 82.7%, 55.7%, 52.2%, and the average is 51.4% (as depicted in Fig. 8B). This result indicates that more than half (51.4%) of the errors were due to maintaining the original solution steps.

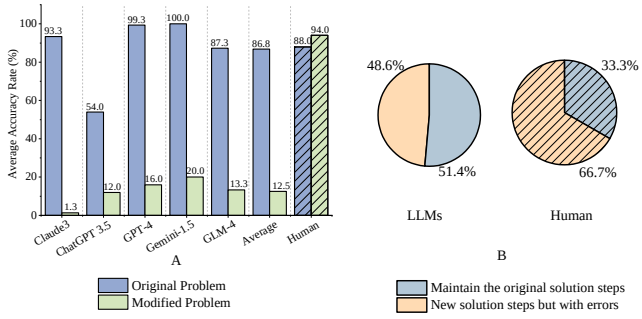


Figure 8: Accuracy of the LLMs and Human(with slash) when answering the CRT3 problems (A), and the proportion of incorrect answer types in modified CRT3 problems (B).

Result Analysis of Experiment II: The modifications to CRT1, CRT2, and CRT3 altered the mathematical principles underlying these problems to preserve the similarity of their statements. The experimental findings reveal that LLMs sometimes rely on their inherent problem-solving methodologies, even when provided with CoT prompts as guidance. This observation further supports our hypothesis that LLMs can address mathematical problems, likely due to the inclusion of analogous problems in their training datasets. Consequently, LLMs select solutions based on the superficial similarity of problem descriptions rather than demonstrating an understanding of the underlying mathematical principles.

Experiment III: Replicating Experiment I & II on the Latest o1 Model

Given OpenAI's release of the o1-preview model, which focuses on enhancing logical reasoning capabilities(OpenAI, 2024), we have repeated the above experiments with it. The results are shown in Fig. 9.

In terms of replicating Experiment I (Fig. 9A), changing the numbers in the problem statements did not significantly affect the accuracy of o1's answers. This observation suggests that the specific number given in the problem statement does not affect the method that o1 selects for solving the problem. This might imply that o1 has potentially incorporated prompts like "list the equations before solving" into its built-in thought process. Nevertheless, this speculation cannot be officially confirmed, as OpenAI has not released technical de-

tails regarding o1's improved reasoning capabilities.

However, when replicating Experiment II (Fig. 9B), the average accuracy of o1 was only 10.0%. After the mathematical principles of the CRT problems are altered while modifying the textual description as little as possible, o1 persisted in selecting problem-solving approaches based on the mathematical principles corresponding to the original problems. The errors stemming from this persistence constituted 100%, 68.4%, and 75.7% of the incorrect answers to the three types of CRT problems, respectively.

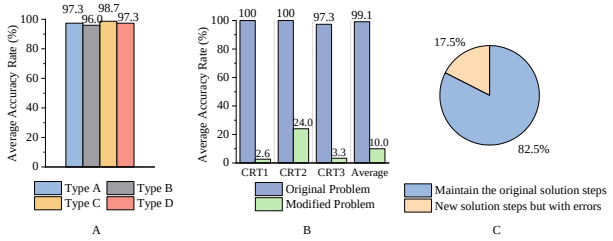


Figure 9: Average Accuracy of the o1 model in Experiment I (A) and Experiment II (B), and the average proportion of incorrect answers in Experiment II (C).

As shown in Fig. 9C, on average, 82.5% of the incorrect answers were due to treating the modified problem as the original one and providing the same answer as for the original problem. This may indicate that technologies such as CoT, prompt engineering, and even fine-tuning tailored to specific datasets (e.g., the o1 model) cannot fundamentally enhance the ability of LLMs to comprehend mathematical problems. The reason for this lies in the fact that the learning paradigm of LLMs has not undergone substantial changes, such as the adoption of auto-regressive next-token predictors, resulting in their thought patterns being deeply ingrained, akin to human intuition (System 1) rather than logical reasoning (System 2).

Related Works

Some previous empirical research has questioned the reasoning abilities of LLMs(Mirzadeh et al., 2024; Sprague et al., 2024; Wu et al., 2024; Zhang et al., 2024), particularly in solving mathematical problems. In contrast, the contributions of this paper are more pronounced in the following aspects:

Most of the previous related studies were published before the release of the o1 model (e.g.Sprague et al., 2024; Wu et al., 2024; Zhang et al., 2024), therefore did not assess this model, which is renowned for its reasoning abilities. The latest experimental evidence provided in this paper demonstrates that similar conclusions remain valid for the newly released o1 model.

Although some previous studies have found that some LLMs exhibit overfitting and memorization effects during the reasoning process, leading to a decline in performance on modified datasets, the degree of decline was generally not significant (e.g.Zhang et al., 2024 , less than 13%), and not all LLMs experienced a decrease. Consequently, these studies were cautious in drawing conclusions, given that even for human testers, a moderate decline is inevitable. However, in this paper, the accuracy of all LLMs drops by more than 50% on average, providing more convincing experimental evidence. We utilized System 1 and System 2 theories from cognitive psychology to examine the incorrect responses produced by LLMs. Our analysis revealed a notable decrease in the accuracy of these responses, which contrasts sharply with the performance of human System 2.

Some previous studies have also mentioned human cognitive science to compare and analyze the thinking patterns of LLMs(Mirzadeh et al., 2024; Wu et al., 2024.) However, their datasets used for testing mathematical reasoning ability are purely mathematical and unrelated to cognitive science itself. In contrast, our dataset itself is a mathematical test dataset from cognitive science, originally used to study the roles of human reasoning (System 2) and intuition (System 1). Therefore, this paper offers greater interpretability and inspiration from a psychological perspective.

Conclusion

This paper draws on the classic CRT problems from human psychology to conduct an empirical study on the "emergence" of mathematical capabilities in mainstream LLMs. It aims to test whether LLMs possess mathematical reasoning capabilities similar to human System 2 and to provide more interpretable explanations for the causes from the perspective of human psychology. By constructing forward experiments (Experiment I) and reverse experiments (Experiment II), we obtained conclusions that are starkly different from mainstream views. Specifically:

- LLMs tend to match problem-solving strategies based on textual similarity rather than truly understanding the underlying principles of mathematical problems. This process is more akin to human intuition (System 1) rather than logical reasoning (System 2).
- Even with the introduction of CoT or specialized training to enhance reasoning abilities (e.g., o1), such methods cannot fundamentally alter the problem-solving thinking patterns of LLMs to endow them with System 2-like rational logical reasoning abilities. The emergence of this phenomenon is perhaps attributable to the dominant paradigm utilized in the training and fine-tuning of LLMs, which fundamentally involves predicting the next token with the highest probability. This mechanism closely aligns with human intuition (System 1), which enables rapid decision-making based on patterns correlated with high probability, thereby presenting a certain distinction from System 2.

This study conducts an empirical analysis of the "emergence" of LLMs' mathematical reasoning abilities from a psychological perspective. We hope that this study can reduce overblown expectations of LLMs' capabilities and stimulate more empirical research to objectively evaluate the current limitations of LLMs' abilities.

References

Anthropic. (2023). *Introducing claude* [Accessed: 2025-05-10]. https://www.anthropic.com/index/introducing-claude

Azim, P. S. H. M. (2024). A Framework for LLM-Assisted Smart Policing System. *IEEE Access*, *12*, 74915–74929.

Bi, Z., Hajialigol, D., Sun, Z., Hao, J., & Wang, X. (2024). STOC-TOT: Stochastic Tree-of-Thought with Constrained Decoding for Complex Reasoning in Multi-Hop Question Answering [arXiv:2407.03687 [cs]].

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., … Amodei, D. (2020). Language models are few-shot learners [arXiv:2005.14165 [cs]].

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4 [arXiv:2303.12712 [cs]].

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2022). GLM: General language model pretraining with autoregressive blank infilling. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Gemini Team, Anil, R., Borgeaud, S., & Alayrac, J.-B. (2024, June). Gemini: A Family of Highly Capable Multimodal Models [arXiv:2312.11805 [cs]].

Goyal, S., Rastogi, E., Rajagopal, S. P., Yuan, D., Zhao, F., Chintagunta, J., Naik, G., & Ward, J. (2024). Healai: A healthcare llm for effective medical documentation. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining(WSDM 2024)*.

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*(10), 833–838.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus; Giroux.

Li, J., Dada, A., Puladi, B., Kleesiek, J., & Egger, J. (2024). Chatgpt in healthcare: A taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, *245*, 108013.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models [arXiv:2410.05229 [cs]].

OpenAI. (2024). *Openai o1 system card* [Accessed: 2025-05-10]. https://cdn.openai.com/o1-system-card.pdf

OpenAl. (2023). *Introducing chatgpt* [Accessed: 2025-05-10]. https://openai.com/blog/chatgpt

Pérez-Núñez, A. (2023). Exploring the potential of generative ai (chatgpt) for foreign language instruction: Applications and challenges. *Hispania*, *106*, 355–362.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? [arXiv:2304.15004 [cs]].

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Sprague, Z., Yin, F., Rodriguez, J. D., Jiang, D., Wadhwa, M., Singhal, P., Zhao, X., Ye, X., Mahowald, K., & Durrett, G. (2024). To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning [arXiv:2409.12183 [cs]].

Stanovich, K. (1999). Who Is Rational?: Studies of individual Differences in Reasoning.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *The 11th International Conference on Learning Representations (ICLR 2023)*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [arXiv:2201.11903 [cs]].

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., & Kim, Y. (2024). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL-HLT 2024)*, 1819–1862.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *The 37th International Conference on Neural Information Processing Systems(NeurIPS 2023)*.

Yu, Z., Zhang, X., Shang, N., Huang, Y., Xu, C., Zhao, Y., Hu, W., & Yin, Q. (2024). WaveCoder: Widespread And Versatile Enhancement For Code Large Language Models By Instruction Tuning [arXiv:2312.14187 [cs]].

Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Slack, D., Lyu, Q., Hendryx, S., Kaplan, R., Lunati, M., & Yue, S. (2024). A Careful Examination of Large Language Model Performance on Grade School Arithmetic [arXiv:2405.00332 [cs]].