# Understanding the Limitations of Mathematical Reasoning in Large Language Models

Dzhos Oleksii
Faculty of Informatics, NaUKMA

29 October 2024

**Abstract**

This paper reviews a scientific publication [1] that demonstrates how large-scale language models lack the ability to perform true logical reasoning, relying primarily on learned data patterns. The study was conducted using the GSM8K benchmark, which revealed significant limitations in the models' mathematical reasoning abilities.

## 1 Introduction

Large language models (LLMs) are increasingly used in various aspects of everyday life, but their responses raise questions about the reasoning processes behind their outputs. Research shows that small changes in the input data—such as token sequences—can dramatically alter the model's output, indicating that the data is highly sensitive. The GSM8K dataset is a popular benchmark for evaluating the mathematical abilities of LLMs. However, it has limitations, as it provides only a fixed set of questions, restricting the assessment of deeper mathematical reasoning. Key points of this proposed improvement include:

- The new GSM-Symbolic benchmark introduces a variety of question formats using symbolic tables, as shown in Figure 1.
- The performance of all models drops on GSM-Symbolic, highlighting unjustified variance in different versions of the same question.
- LLMs are not sensitive to names but show sensitivity to numerical values. Increased degradation and variance in clauses suggest compounding issues in LLM reasoning.
- The primary problem with modern LLMs is that their reasoning is not formal, but based on template matching.
- GSM-NoOp is introduced, containing irrelevant but contextually related information, which helps uncover deeper reasoning issues that cannot be addressed by simple contextual adjustments.

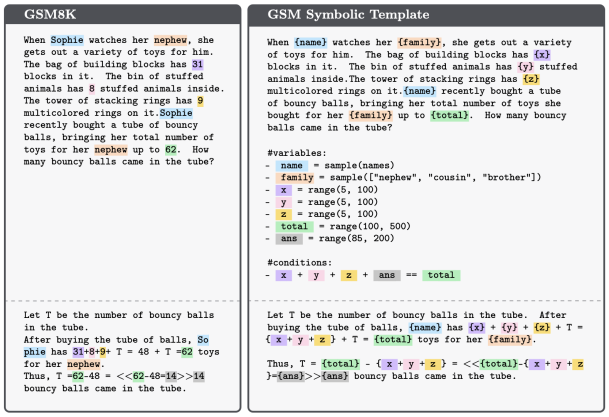


Figure 1: GSM-Symbolic

## 2 Reasoning and Language Models

The main difference between LLMs and true intelligent systems is their capacity for logical reasoning. Studies suggest that complex reasoning in LLMs requires additional computation and data structures, like attention and feedback mechanisms, also known as Chain of Thought (CoT). CoT helps process tokens across different levels, with each level corresponding to a specific moment in the reasoning process.

LLMs do not "think" in the traditional sense; instead, they search through templates, which leads to problems such as inconsistent confidence in outputs, where small changes can result in different outcomes. This decreases the sensitivity of results when performing similar tasks, as neighboring classifiers in CoT are similar, thus limiting LLM reasoning. Another issue is that the accuracy of responses decreases as tasks involve more tokens, and the precision of reasoning correlates with the frequency of patterns in training data.

The introduction of GSMS marks a new stage in LLM evaluation, with templates for creating question pools that provide a detailed assessment of model reasoning. Scores across various tests indicate that LLMs lack formal mathematical reasoning, emphasizing the need for continued research to enhance their reasoning capabilities.

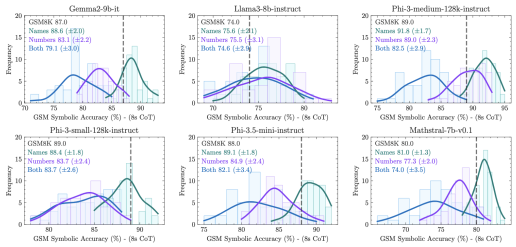


Figure 2: Sensitivity LLM

# 3 GSM-Symbolic

The GSM8K dataset contains 7,473 questions and 1,319 test items. As shown in Figure 1, these questions are easy for humans to understand, but due to the presence of redundant information, slight changes can significantly affect results. We created the iGSM dataset, which captures parameter dependencies in a hierarchical, graphical structure. In scientific publishing, LLMs face difficulty when answering the same question repeatedly, revealing fundamental problem-solving deficiencies. Therefore, problem-solving cannot be addressed merely through hints or by tweaking difficulty levels.

In this paper, we evaluate 20 LLMs, ranging in size from 2B to 27B parameters. While existing benchmarks offer a performance metric for a fixed set of questions, this work proposes evaluating LLMs across a distribution of problem instances. We conducted 500 evaluations under varying conditions—100 templates, with 50 examples each—resulting in 5,000 examples per benchmark. Each example is a mutation of one of the 100 GSMK examples, using the CoT method.

# 4 Experiments and Results

In Figure 2, the results of our experiments demonstrate a consistent decline in LLM performance when using the GSMS benchmark. Initially, the performance on GSM8K remains stable when only the names in the test cases are altered. However, a noticeable drop occurs when numerical values are changed, and the decline becomes even more significant when both names and values are modified simultaneously. Additionally, as the number of sentences in the test cases increases, the average performance score decreases, while the variance of the results consistently widens across all models.

The first experiment is to evaluate the performance of the models using the GSMS test. Figure 2 illustrates the distribution of performance across the models, all of which have non-zero variance. This is surprising, given that the differences between the versions of the questions are only in the names and numerical values, while the overall reasoning steps remain identical. Another noteworthy observation is the deviation from the center of the distribution, which indicates that certain models perform inconsistently even with seemingly minor changes in the input.

The second experiment assesses the fragility of reasoning in the LLM. The models attempt to perform matching by matching the questions asked and the solution steps with those seen during training, resulting in high variance. The experiment involves changing proper names and numerical variables. Initially, the accuracy of the GSM8K models more closely matches the distribution when only proper names are changed compared to changes in numerical values, but as the complexity of these changes increases, performance decreases and variance increases, which generally highlights the weakness of the logical abilities of modern LLMs.

The third experiment evaluates the complexity of variance in academic performance distribution by generating new GSM-Symb templates and modifying them. Removing an element created GSM-Symbolic-Minus-1 (GSM-M1), while increasing difficulty by adding sentences resulted in GSM-Symbolic-Plus-1 (GSM-P1) and GSM-Symbolic-Plus-2 (GSM-P2). The results show that as question complexity increases, performance declines and variance rises, indicating that the models are not engaging in formal reasoning. Although the number of reasoning steps increases linearly, the accuracy decline accelerates disproportionately.

# 5 Conclusion

This research paper examines how LLMs reason and highlights limitations in current performance estimates using the GSM8K benchmark. It introduces the GSM-Symbolic benchmark for broader reasoning tests and finds significant variance in model performance across question variants, raising concerns about GSM8K's reliability. The study reveals fragile reasoning in LLMs, with accuracy dropping as difficulty increases and sensitivity to irrelevant information. Both benchmarks focus on simple math, suggesting that these limitations would be more pronounced in complex tasks. Evaluating future models remains critical for developing systems with true cognitive reasoning abilities.

# References

[1] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models," *Apple*, 2024. [Online]. Available: https://arxiv.org/pdf/2410.05229