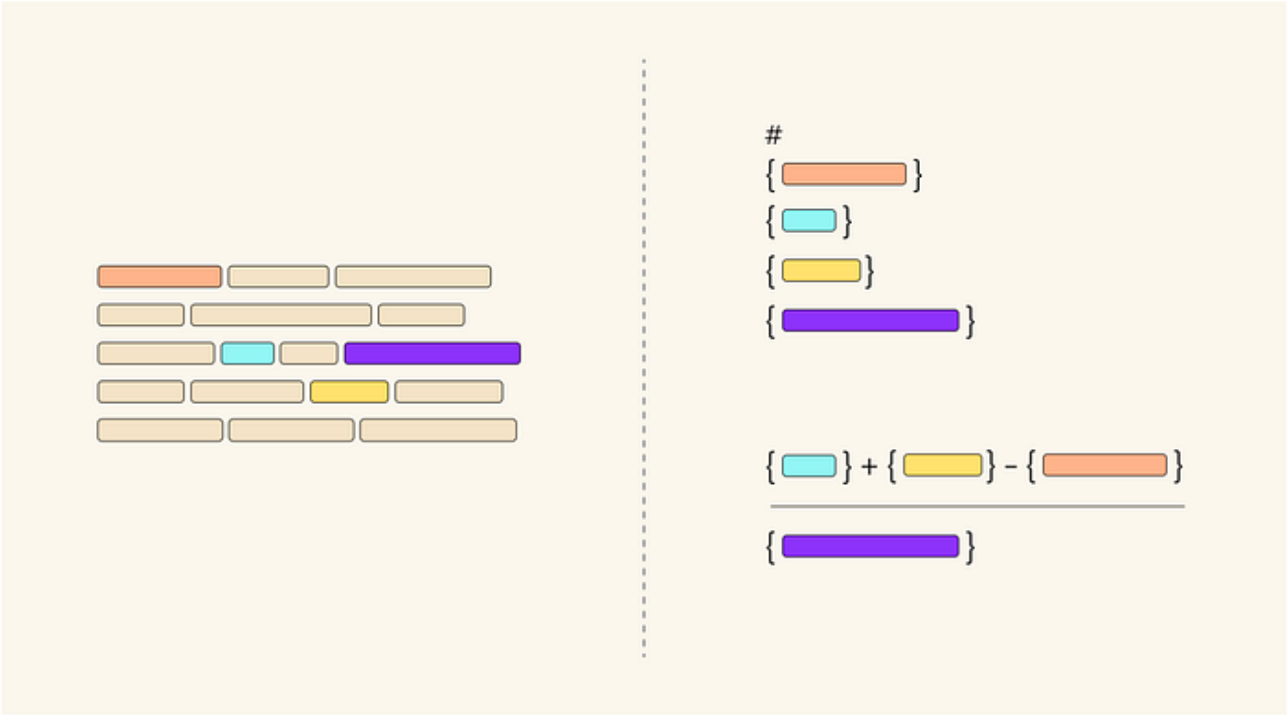
MEDIUM – 28 OTTOBRE 2024

GSM-Symbolic: Analyzing LLM Limitations in Mathematical Reasoning and Potential Solutions

di Alexander Watson



Source: Gretel.ai

Introduction

Large language models (LLMs) have recently made significant strides in AI reasoning, including mathematical problem-solving. However, a recent paper titled "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models" by Mirzadeh et al. raises questions about the true capabilities of these models when it comes to mathematical reasoning. We have reviewed the paper and found it to be a valuable contribution to the ongoing discussion about AI capabilities and limitations, however, our analysis suggests that its conclusions may not fully capture the complexity of the issue.

The GSM-Symbolic Benchmark

The authors introduce GSM-Symbolic, an enhanced benchmark derived from the popular GSM8K dataset. This new benchmark allows for the generation of diverse question variants, enabling a more nuanced evaluation of LLMs' performance across various setups. The study's large-scale analysis of 25 state-of-the-art open and closed models provides significant insights into how these models behave when faced with mathematical reasoning tasks.

Press enter or click to view image in full size

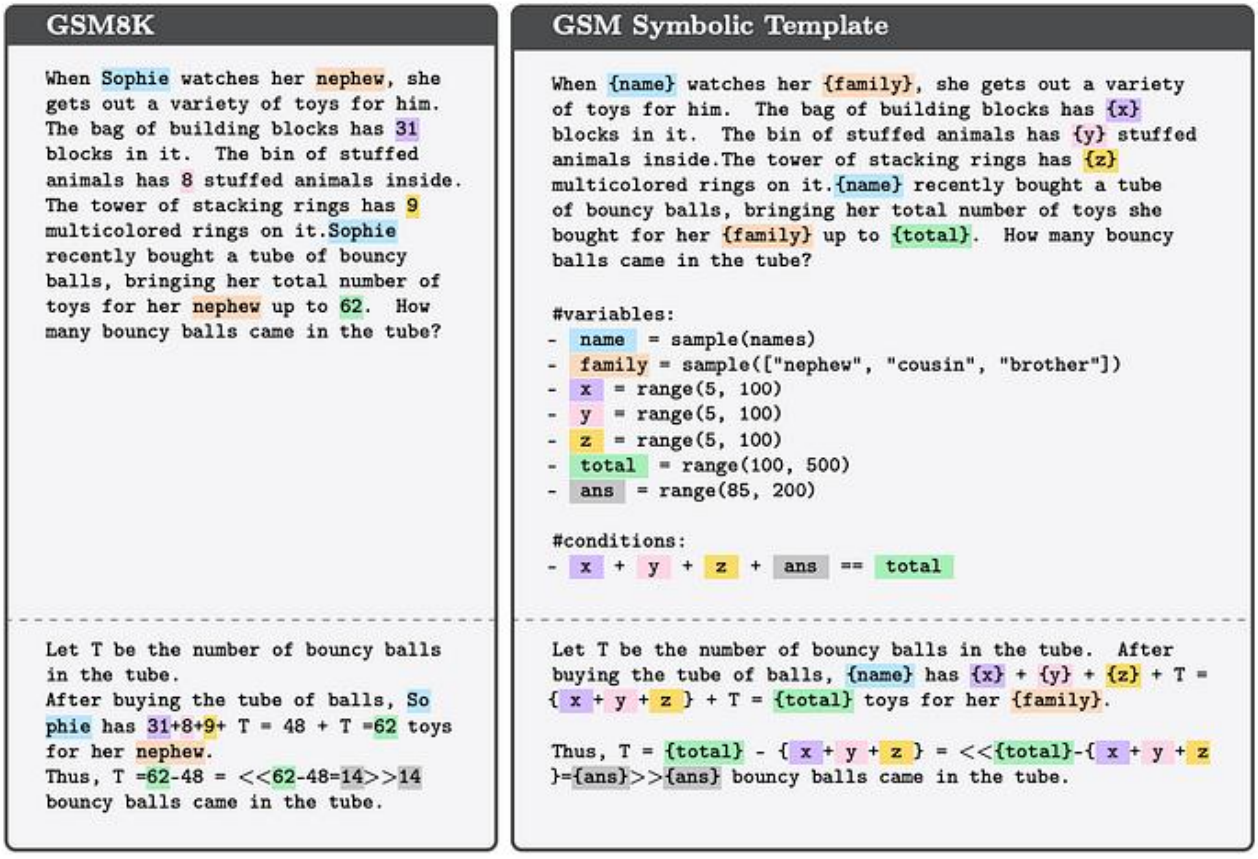


Figure 1: GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models (Source: Mirzadeh et al., GSM-Symbolic Paper)

Performance Variability and Model Comparisons

One of the most surprising findings is the high variability in model performance across different instantiations of the same question. All models exhibit "significant variability in accuracy" when tested on GSM-Symbolic. This variability raises concerns about the

reliability of currently reported metrics on the GSM8K benchmark, which relies on single point-accuracy responses.

Press enter or click to view image in full size

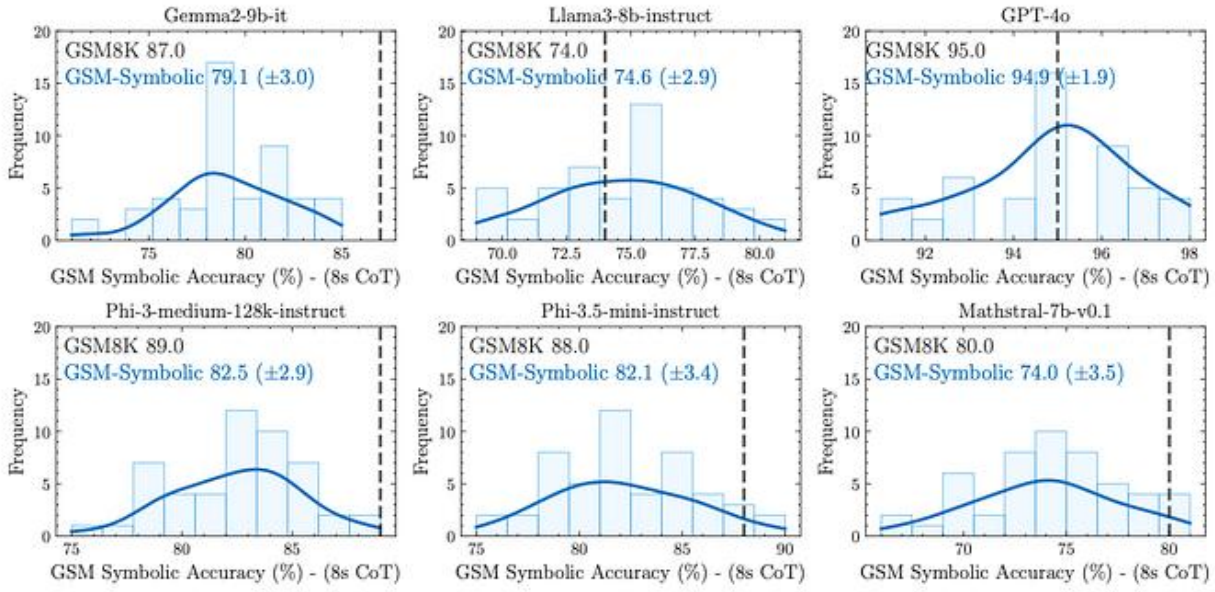


Figure 3: GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models (Source: Mirzadeh et al., GSM-Symbolic Paper)

Not all models are created equal. Llama-3–8b and GPT-4o are clear outliers in that they don't exhibit as significant of a drop on the new benchmark as other models like gemma-2–9b, phi-3, phi-3.5 and mathstral-7b. This observations suggests two important points:

1. Llama-3–8b and GPT-4o generally demonstrate a more robust understanding of mathematical concepts, although they are still not immune to performance variations.
2. The training data for Llama-3–8b and GPT-4o likely has not been contaminated (or at least not to the same extent) with GSM8K data. In this context, data contamination refers to the unintentional inclusion of test or benchmark data in a model's training set, leading to artificially inflated model performance during evaluation. If contamination had occurred, as the authors hypothesize for some models, we would expect to see very high performance on GSM8K

but significantly lower performance on even slight variations of these problems.

These findings highlight a opportunity for improvement through the use of synthetic data, where properly designed synthetic datasets can address both of these points for anyone training models:

1. To mitigate potential data contamination issues, there's no need to use the original GSM8K data in training when high-quality synthetic versions can be generated (blog link). These synthetic datasets retain the mathematical reasoning challenges of GSM8K without reusing the exact problems or solutions, thus preserving the integrity of the model's evaluation.
2. Even more importantly, it's possible to generate synthetic data that surpass the quality of both the OpenAI GSM8K and Apple GSM-Symbolic datasets. This approach can lead to a more robust understanding of mathematical concepts, addressing the performance variability observed in current models.

Sensitivity to Changes and Complexity

The authors show that LLMs are more sensitive to changes in numerical values than to changes in proper names within problems, suggesting that the models' understanding of the underlying mathematical concepts may not be as robust as previously thought. As the complexity of questions increases (measured by the number of clauses), the performance of all models degrades, and the variance in their performance increases. This highlights the importance of using diverse data in training, and this is something that synthetics can help with. As the authors demonstrate, there is logically no reason why a AI model should perform worse on a given set of problems, with just a simple change in numbers or a slight variation in the number of clauses.

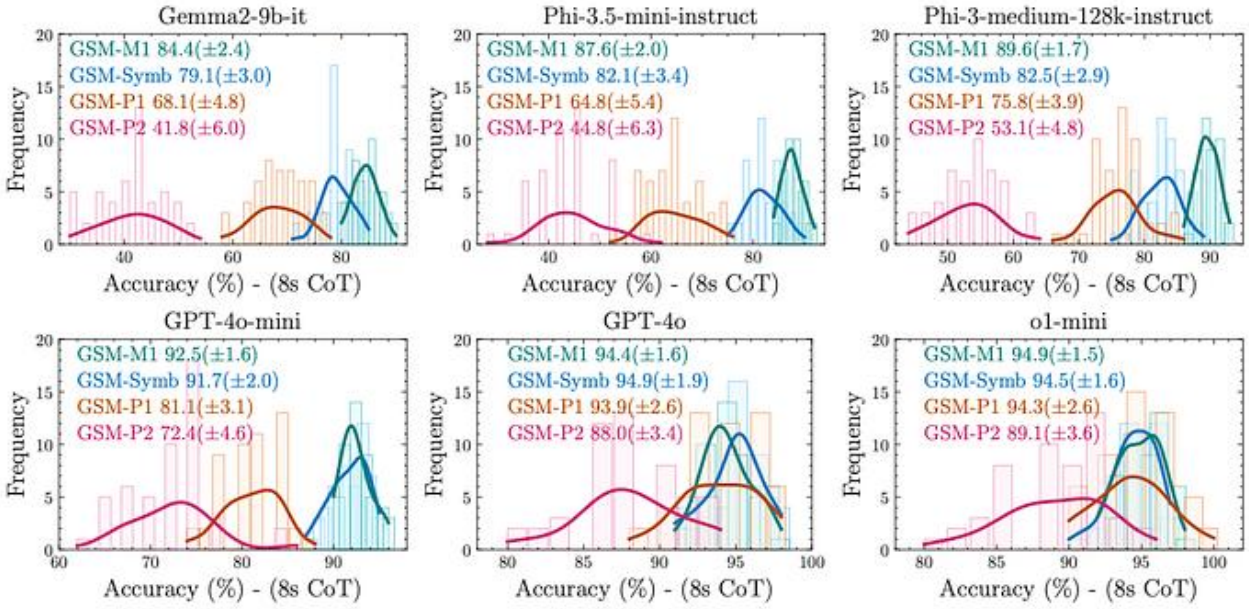


Figure 4: GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models (Source: Mirzadeh et al., GSM-Symbolic Paper)

The GSM-NoOp Challenge

Perhaps the most concerning finding is the introduction of GSM-NoOp, a dataset designed to challenge the reasoning capabilities of LLMs. By adding seemingly relevant but ultimately inconsequential information to problems, the authors observed substantial performance drops across all models — up to 65% for some. The authors propose that this points to current LLMs relying more on a type of pattern matching than true logical reasoning

Press enter or click to view image in full size

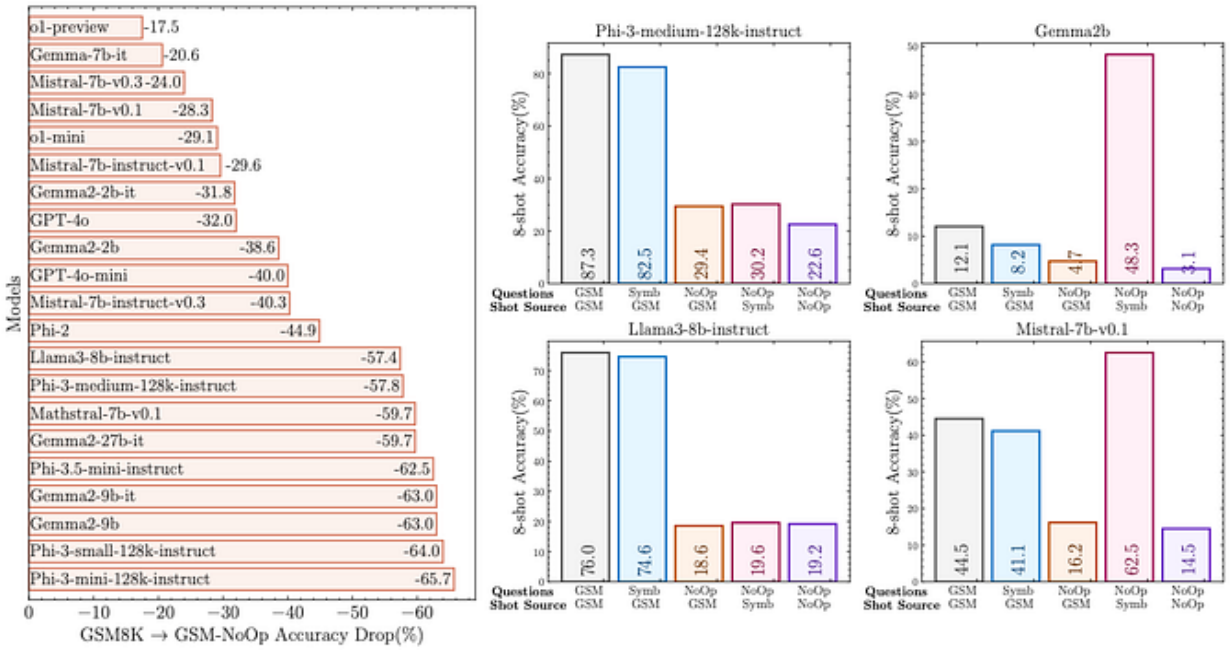


Figure 6: GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models (Source: [Mirzadeh et al., GSM-Symbolic Paper])

A Critical Perspective on the Paper's Conclusions

While the GSM-Symbolic study provides valuable insights into the performance of LLMs on mathematical reasoning tasks, it's important to critically examine the paper's conclusions. The authors argue that the observed limitations suggest LLMs are not capable of true logical reasoning. However, this interpretation may be oversimplifying a complex issue.

The paper's argument for LLMs relying on pattern matching rather than reasoning seems less definitive when examined closely. It's clear that these models are not perfect reasoners — if they were, they would achieve 100% accuracy on GSM8K. But the leap from imperfect performance to a lack of reasoning capability is not necessarily justified.

There are at least two potential explanations for why LLMs, like humans, sometimes get questions wrong:

1. The model tries to strictly pattern match a problem to something it has seen before, and fails if it can't.

2. The model tries to follow a logical program but has a certain (compounding) probability of making an error at each step, as expected based on the fact that it literally samples tokens.

The paper seems to lean towards explanation (1), but doesn't make a convincing case for why this should be preferred over explanation (2). In fact, (2) is more akin to human-like reasoning and potentially more interesting from a research perspective.

Let's examine each main finding of the paper through this critical lens:

GSM-Symbolic Performance

The GSM-Symbolic approach is a valuable method for dataset expansion, validating the potential of synthetic data generation techniques like those used by Gretel. However, it's worth noting that model performance doesn't completely fall apart on these new variants — it just gets somewhat worse. If the models were strictly pattern matching, we might expect performance to drop to near zero on these new variants. The observed behavior seems more consistent with a model that can generalize to some degree but makes more errors on unfamiliar problem structures.

Even human experts are not infallible. On the MATH benchmark, for instance, former math olympians typically scored 18/20 or 19/20, making small arithmetic errors. This suggests that error-prone reasoning, rather than a lack of reasoning capability, might be a more accurate description of both human and LLM performance.

Varying Difficulty

The paper's findings on performance degradation with increasing question complexity are consistent with the idea of compounding errors in a multi-step reasoning process. As the number of steps increases, so does the probability of making an error at some point in the chain. This behavior is observed in human problem-solving as well and doesn't necessarily indicate a lack of reasoning ability.

GSM-NoOp Challenge

The GSM-NoOp results, may not be as directly related to reasoning capability as the paper suggests. In real-world scenarios, we typically assume that all information provided in a problem statement is relevant. For instance, in the example question in

Figure 7, a reasonable human might infer (like the LLMs did) that the size of the kiwis was only mentioned because they were discarded.

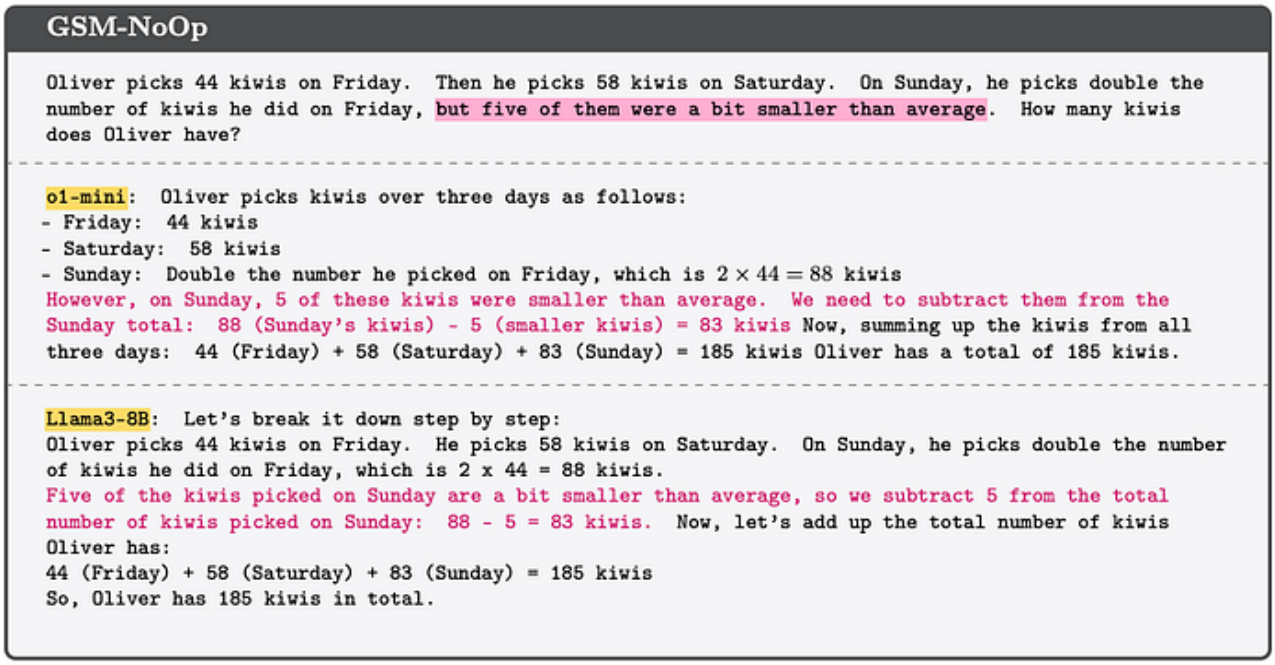


Figure 7: GSM-Symbolic: Example GSM No-Op question. (Source: Mirzadeh et al., GSM-Symbolic Paper)

The ability to discern relevant information from irrelevant information, especially when the irrelevant information is inserted with the intent to be misleading (i.e. *seemingly* relevant), is a separate skill from pure mathematical reasoning.

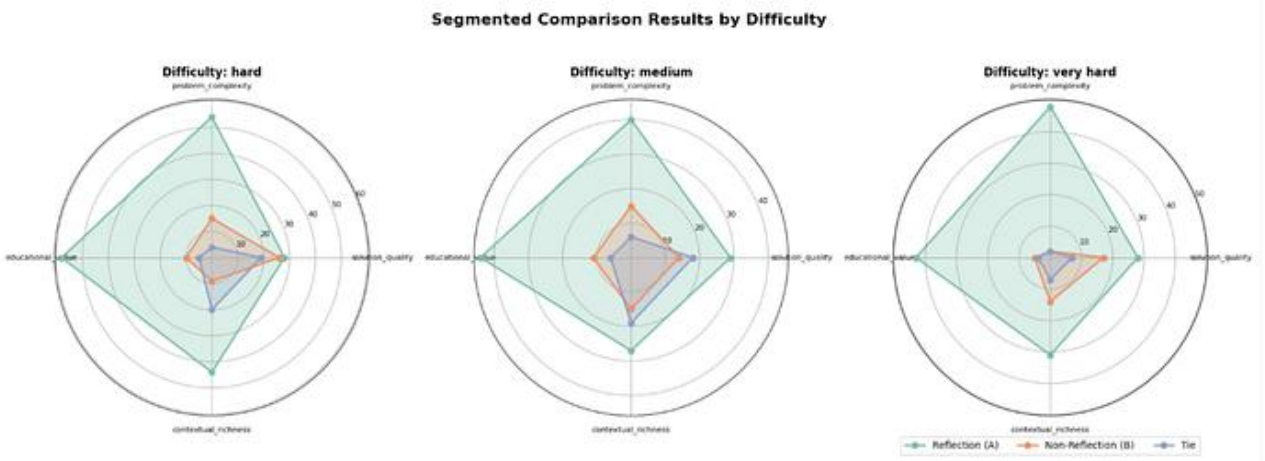
The authors include a follow-up experiment (NoOp-NoOp) in which the models are implicitly "warned" of the misleading intent: they use few-shot examples that also contain irrelevant information. The subset of models illustrated with this experiment still show a drop in performance. **Several follow-up experiments could serve to better understand the phenomenon**:

1. Expand the NoOp-NoOp experiment to more models;
2. Measure how well models perform when *explicitly* warned that some information may be irrelevant in the prompt;
3. Fine-tune models on synthetic training examples that include irrelevant information in addition to examples that contain entirely relevant information.

Opportunities for Improvement: The Promise of Synthetic Data

While the paper by Mirzadeh et al. highlights important limitations in current LLMs, at Gretel we have developed datasets that address many of the challenges identified in the paper:

1. **Synthetic GSM8K** Dataset: Available on HuggingFace at [gretelai/synthetic-gsm8k-reflection-405b](), this dataset focuses on generating more complex, multi-step reasoning versions of problems than what existed in the original human generated dataset from OpenAI. It incorporates advanced prompting techniques, including Reflection and other cognitive models, to capture detailed reasoning processes. This approach has shown significant improvements, particularly for very hard problems, demonstrating its potential to enhance AI's ability to handle complex, multi-step reasoning tasks. As covered in our blog, Gretel's synthetic data created using these techniques achieved a [92.3% win-rate on problem complexity and an 82.7% win-rate for educational value over the standard Llama 3.1 405B parameter model outputs](), using these advanced techniques as judged by GPT-4o- demonstrating that LLM reasoning can further be unlocked with more sophisticated training data examples and prompting techniques than the basic Chain-of-Thought used in the paper.



Source: [https://gretel.ai/blog/teaching-ai-to-think-a-new-approach-with-synthetic-data-and-reflection]()

2. Synthetic Text-to-SQL Dataset: Generated by Gretel to help improve LLMs ability to interact with SQL-based databases/warehouses & lakes, available at gretelai/synthetic_text_to_sql, has proven highly effective in improving model performance on Text-to-SQL tasks. When used to fine-tune CodeLlama models, it led to 36%+ improvements on the BIRD benchmark, a challenging cross-domain Text-to-SQL evaluation platform. Further supporting the theory about today's LLMs being trained on data that is too simple and leading to memorization, a single epoch of fine-tuning the Phi-3 and Llama 3.1 models on this dataset yielded a 300%+ improvement on BIRD benchmark problems labeled as "very hard".

These results demonstratethat high-quality synthetic data can be a powerful tool in addressing the limitations of current LLMs in complex reasoning tasks.

Future Directions

In conclusion, the GSM-Symbolic paper provides valuable insights into the current limitations of LLMs in mathematical reasoning tasks. However, its conclusions should be approached critically. The observed behavior of LLMs could be interpreted in multiple ways, and it's possible that the paper's emphasis on pattern matching over reasoning may be oversimplifying a more complex issue.

The limitations identified by the study are real and significant. The variability in performance, sensitivity to numerical changes, and struggles with irrelevant information all point to areas where current LLMs can be improved.

However, as demonstrated by more advanced models such as GPT-4o and Llama 3.1 above- by synthesizing diverse, challenging problem sets that push the boundaries of what AI models can tackle, we can develop LLMs that exhibit more robust, human-like reasoning capabilities.

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. 2024.