Can LLMs *understand* Math? Exploring the Pitfalls in Mathematical Reasoning

Aditeya Baral
Courant Institute of Mathematical Sciences
New York University
ab12057@nyu.edu

Ayush Rajesh Jhaveri
Courant Institute of Mathematical Sciences
New York University
aj4332@nyu.edu

Tiasa Singha Roy
Courant Institute of Mathematical Sciences
New York University
ts5478@nyu.edu

Yusuf Baig
Courant Institute of Mathematical Sciences
New York University
yb2510@nyu.edu

1 Overview

Motivation. Large Language Models (LLMs) have shown impressive capabilities across tasks such as text generation, language translation, question answering, and sentiment analysis. However, their performance diminishes when it comes to complex reasoning tasks, particularly in mathematical domains. While LLMs tend to perform well on elementary math problems, they often struggle with complex mathematical reasoning, leading to errors in tasks involving precise, step-by-step problem-solving. Moreover, in multi-step tasks, where sequential reasoning is crucial, LLMs frequently make errors, failing to arrive at correct solutions. This highlights the need for a holistic evaluation of their mathematical reasoning abilities to identify these limitations and develop targeted improvements.

Related work. We draw on insights from ReasonEval[1], which argues that solely relying on final answer accuracy can mask the use of unnecessary or incorrect intermediate steps in the mathematical reasoning process. It introduces a methodology that highlights the importance of going beyond accuracy in evaluating LLM performance for mathematical reasoning. To extend this methodology, we leverage ideas of self-reflection[2], which proposes a method for using LLMs to self-correct themselves for reasoning. We take motivation from this method to use the LLM itself for identifying pitfalls and patterns in its reasoning evaluation. However, these methods depend on external sources for effective self-improvement. Our work builds on this by using oracle labels directly within the LLM, allowing it to autonomously identify and analyze patterns in its reasoning failures. Furthermore, while past studies, such as [3][4], argue that using oracle labels for self-correction may not be realistic for all applications, we propose employing them here in a self-feedback context.

Goal. Prior work [1] observes how current LLMs and evaluation methods lack in reasoning tasks, particularly complex mathematical reasoning. We hope to extend this work and bridge any reasoning gaps with our work by leveraging self-reflection techniques to understand potential issues in reasoning and create an evaluation framework that accounts for such issues. We also hope to explore generalized ideas in error patterns within the identified issues and how they can be addressed in the initial question-answering process to improve accuracy.

2 Project plan

Methods. We plan to expand on the work done in [1] by identifying key issues in mathematical reasoning by using a structured approach involving data selection, model evaluation and analysis prompting strategies. We summarize the steps as follows:

1. **Dataset**: We assess LLMs' consistency and logical fallacies in mathematical reasoning using a variety of math word problem (MWP) datasets, ranging from primary school to undergraduate-level complexity. In addition to GSM8K [5] used in [1], we also assess LLMs with TAL-SCQ5K-EN[6], and MATH to provide a dataset-independent study. We analyze trends in reasoning strength and identify common pitfalls as LLMs tackle problems specific to particular mathematical topics and higher-order difficulty levels, an aspect which has not been covered in prior work.
2. **Evaluation of mathematical datasets with step-wise ground truth solutions**: We will assess various LLMs on solving math word problems by measuring their solving accuracy, reasoning capability (accuracy of step-by-step logic using frameworks like ROSCOE, ReasonEval and LLMReasoners), and self-correction ability. Using a wider range of frameworks, we hope to cover all potential error cases across the datasets mentioned above.
3. **Prompting techniques**: We compare different prompting strategies to understand LLMs' mathematical reasoning in failing cases after evaluation. This is done by providing the model with ground truth answers to leverage self-reflection ideas to prompt the models in identifying errors. This is the key difference in our approach compared to the previous work done in [1]. By utilizing this, we can identify potential error patterns in reasoning and we can not only extend the evaluation framework but also explore initial question-answering prompts to improve the process.

Baselines. Our work does not need a distinctive baseline since we focus on analyzing flaws in mathematical reasoning.

Data. We evaluate LLMs' consistency and logical fallacies in mathematical reasoning by benchmarking them against standardised math word problem (MWP) solving datasets that vary in linguistic and mathematical complexity. We additionally explore datasets that cover primary school mathematics involving simple operations to high school and undergraduate level mathematics involving trigonometry and calculus. We compare across a range of difficulty levels to analyze and establish trends in the strength of reasoning as these models attempt to solve MWPs of higher-order complexities and identify the difficulties, pitfalls and patterns that might emerge during solving. Since our study focuses on comparing the reasoning ability as well as the final answer, we pick only those datasets which include step-by-step solutions for CoT fine-tuning.

The datasets in increasing order of linguistic and mathematical complexity are listed below.

1. **Grade School Math (GSM) 8K**[5]. The GSM8K dataset comprises 8500 linguistically diverse MWPs from elementary and grade school mathematics. The MWPs can be solved in 2 to 8 steps using basic arithmetic operations (+, -, $\times$ and $\div$) and is one of the simpler datasets chosen for evaluation.
2. **TAL-SCQ5K-EN**[6]. The TAL-SCQ5K-EN dataset comprises 5000 high-quality MWPs of the junior, middle and high school levels. The MWPs are mathematically complex since they have been compiled from various math competitions.
3. **Mathematics Aptitude Test of Heuristics (MATH)**[7]. The MATH dataset comprises 12500 MWPs from various high-school and higher-level math competitions like AMC 10, AMC 12 and so on, thus making it one of the most challenging datasets since it includes topics like calculus, geometry and number theory.

Evaluation. We will investigate the mathematical reasoning ability of multiple LLMs from the GPT, Llama, Mistral and Gemma families to evaluate their performance at understanding and solving MWPs. Our evaluation will focus on the correct answer as well as the step-by-step reasoning performed to reach the final answer. We are specifically concerned with the following characteristics,

1. **Solving Accuracy**. We determine this as the percentage of correct solutions provided, regardless of the steps produced to reach the solution.
2. **Reasoning Capability**. We determine the correctness of the step-by-step approach to reach a solution. To facilitate our study of reasoning ability, we employ approaches that can evaluate the correctness of the intermediate steps.
   (a) **ROSCOE**[8]. The ROSCOE framework can measure the *semantic alignment* and *similarity* between a generated reasoning step against a ground truth. We aim to leverage this framework to better understand and compare the reasoning ability of our chosen LLMs.
   (b) **ReasonEval**[1]. This baseline is adept at scoring the *validity* and *redundancy* of any intermediate steps while solving MWPs. We aim to benchmark our chosen LLMs against these metrics to numerically establish the overall validity and redundancy of the solving approach.
   (c) **LLMReasoners**[9]. This evaluation framework establishes fixed *criteria* for reasoning chains to generate intermediate steps and can evaluate how strictly LLMs can reason within these constraints. We aim to benchmark our chosen LLMs against these reasoning constraints to evaluate the correctness and relevance of the approach.
3. **Self-Correction Ability**. Given an incorrect solution, we determine the LLM's ability to identify the pitfall in its incorrect approach and correct it if possible to reach the right solution. Additionally, we aim to classify these errors as limited but not exhaustive to arithmetic mistakes, logical inconsistencies, incomplete reasoning, incorrect application of facts or formulas and failure in problem interpretation.

Compute. We aim to evaluate our framework on the GPT-4, Llama 3.1, Mistral 7B (MathStral-7B-v.01) and Gemini 1.5 Pro family of LLMs. We estimate that the total compute requirement will depend on the number of models tested and the evaluation process involved.

References

[1] Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. *arXiv preprint arXiv:2404.05692*, 2024.

[2] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

[4] Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *arXiv preprint arXiv:2406.01297*, 2024.

[5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[6] TAL Education Group. TAL-SCQ5K. `https://huggingface.co/datasets/math-eval/TAL-SCQ5K`, 2023.

[7] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

[8] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*, 2022.

[9] Shibo Hao*, Yi Gu*, Haotian Luo*, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, et al. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*, 2024.