Project Proposal: Exploring the Limitations of Large Language Models in Mathematical Reasoning

Ke Xu
kex005@ucsd.edu

1 Topic and Task

Recent advancements in natural language processing have been spearheaded by the development of large language models (LLMs) such as GPT-3. These models have shown remarkable capabilities across a range of tasks, including text generation, translation, and question-answering. However, their application in mathematical reasoning, particularly in solving math word problems (MWPs), presents unique challenges and limitations. This project aims to explore these limitations, focusing on the capabilities of LLMs in interpreting and solving complex mathematical tasks as systems of equations.

2 Literature Review

Brown et al. (2020) introduced GPT-3, a model demonstrating strong performance across many tasks, including some capabilities in mathematical reasoning. Their study, however, highlighted the model's inconsistency in solving math word problems (MWPs) due to its dependence on the quality and type of training data. They discussed the model's limitations in depth, providing insights into areas where GPT-3 could be improved for enhanced mathematical comprehension. This revelation points towards the need for refined training datasets and potentially adaptive learning algorithms to boost performance. Overall, their research sets a foundational understanding of the capabilities and limitations of current LLMs in mathematical contexts (Brown et al., 2020).

Zong and Krishnamachari (2023) analyzed GPT-3's effectiveness in solving MWPs that can be formulated as systems of linear equations. They found that GPT-3 could classify and generate new problems efficiently, but its ability to extract equations from MWPs accurately required fine-tuning. Their research underscores the importance of targeted model training and fine-tuning in achieving higher accuracy and reliability in mathematical problem-solving. This study provides a pathway for enhancing LLMs' utility in educational and computational mathematics by focusing on their precision and adaptability. Their work is instrumental in advancing our understanding of the practical applications and inherent challenges of using LLMs in complex mathematical scenarios (Zong and Krishnamachari, 2023).

Cobbe et al. (2021) focused on training verifiers to enhance the performance of language models on mathematical problems. They proposed external verification mechanisms as a novel approach to significantly improve the accuracy of mathematical solutions provided by LLMs. Their findings suggest that integrating these mechanisms can address some of the fundamental inconsistencies found in LLM outputs. The study is pivotal as it opens up new avenues for the application of LLMs in fields requiring precise and verifiable computational outputs. This contribution is particularly valuable for developing robust, reliable educational tools that assist in mathematical learning (Cobbe et al., 2021).

Testolin (2023) addresses the evaluation metrics and diversity of datasets in the assessment of LLMs applied to mathematical problems. This work critically examines the current methodologies used in the evaluation of LLMs and identifies a gap in the standardization of metrics and datasets. Testolin argues that this lack of standardization leads to inconsistent results and hinders comparative analysis across different models. He calls for a unified framework that could facilitate more reliable and comparative assessments of LLM capabilities in mathematics. Additionally, he suggests the development of a comprehensive set of new benchmarks that are more reflective of real-world mathematical reasoning challenges

(Testolin, 2023).

Imani et al. (2023) focus on the linguistic capabilities of LLMs and how they can be leveraged to interpret and solve mathematical problems. The paper reviews various techniques for training LLMs specifically for the domain of mathematics, emphasizing the importance of context and semantic understanding. They highlight the challenges LLMs face when transitioning from textual data to mathematical logic, which often involves distinct reasoning patterns. Their research demonstrates that LLMs can perform well on standardized datasets but often fail in real-world scenarios that require adaptive reasoning. The authors propose a set of best practices for dataset preparation and model training to enhance LLMs' performance in mathematical tasks (Imani and colleagues, 2023).

3 Codebase and Task Description

This project employs OpenAI's GPT-3 model to address specific challenges associated with Math Word Problems (MWPs) that involve systems of two linear equations. The codebase, designed to leverage GPT-3's language processing capabilities, focuses on the following tasks:

- **Classifying Problems:** Categorizing problems into predefined themes based on their content, such as 'item and property', 'mixture', 'perimeter of rectangle', and 'sum and difference'.
- **Extracting Equations:** Converting textual descriptions of problems into corresponding systems of equations. For example, for question "In a two digit number. The units digit is thrice the tens digit. If 36 is added to the number, the digits interchange their place. Find the number", the relationships can be modeled by the following equations:

$$x = 3y, \tag{1}$$

$$10y + x + 36 = 10x + y. \tag{2}$$

- **Generating Problems:** Creating new, educative math word problems to aid learning, like from topic 'solid items' and to new topic 'money (funds, bills, stocks, etc.)'.

4 Datasets

The project utilizes a series of custom datasets that range from basic to advanced complexities to train and evaluate the GPT-3 model, from the paper **Solving math word problems concerning systems of equations with gpt-3** (Zong and Krishnamachari, 2023). Datasets include sets of 20, 30, 50, 100, 200, and 1000 problems, each constructed to represent a balance across five predefined categories of MWPs. These problems were sourced from educational websites and books, ensuring a variety of problem types typical of middle and high school levels.

5 Evaluation Metrics

The performance of the GPT-3 model on the MWP tasks is evaluated using the following metrics:

- **Accuracy:** Measured by the model's ability to correctly classify, extract, and generate math word problems.
- **Precision and Recall:** Specifically used for the extraction tasks to evaluate the correctness and completeness of the equations generated by the model.
- **Creativity in Problem Generation:** Assessed by the novelty and variability of the problems generated, ensuring they are logically consistent and solvable.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cobbe, K., Kosaraju, V., Bavarian, M., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Imani, M. and colleagues (2023). Advances in language models for mathematical reasoning. In *Proceedings of the International Conference on Machine Learning*, pages 88–97.

Testolin, A. (2023). Challenges and prospects of using llms in mathematical domains. *Computational Intelligence and Mathematics*, 45(4):659–675.

Zong, M. and Krishnamachari, B. (2023). Solving math word problems concerning systems of equations with gpt-3. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.