Evaluating and Enhancing Large Language Models for Solving College-Level Mathematical Problems Using Retrieval-Augmented Generation and Chain-of-Thought Prompting

Param Somane
Department of Computer Science and Engineering
University of California San Diego
psomane@ucsd.edu

1 Introduction

Mathematical reasoning remains a significant challenge for large language models (LLMs). While recent advances in natural language processing (NLP) have led to impressive capabilities in language understanding and generation, LLMs often struggle with solving complex mathematical problems, particularly at the college level and beyond. Accurate mathematical problem-solving requires not only linguistic competence but also the ability to perform precise calculations and logical reasoning.

Previously, I participated in the *AI Mathematical Olympiad - Progress Prize 1* Kaggle competition (Investments, 2024), which aimed to develop algorithms and models capable of solving challenging math problems written in LaTeX format. Through my participation, I gained valuable experience in applying LLMs to mathematical problem-solving and achieved commendable performance among over a thousand participants. The code I developed during the competition leveraged advanced prompting techniques and model fine-tuning to improve mathematical reasoning capabilities.

This project aims to evaluate and enhance the performance of LLMs in solving college-level mathematical problems by leveraging Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and Chain-of-Thought (CoT) prompting (Wei et al., 2022). I will investigate whether providing LLMs with additional context from mathematical textbooks and lecture notes can improve their reasoning processes and lead to more accurate solutions.

Improving the mathematical problem-solving abilities of LLMs has significant practical implications. It can aid in educational settings, assist in automated theorem proving, and contribute to advancements in fields that rely heavily on mathematical computations. Moreover, understanding the limitations and capabilities of LLMs in mathematical reasoning can inform future research directions in NLP and AI.

2 Approach

To address the challenge of enhancing LLMs' ability to solve college-level mathematical problems, I propose a multi-faceted approach:

1. **Baseline Evaluation**: I will begin by evaluating existing LLMs (e.g., GPT-3, GPT-4, or open-source equivalents) on standard mathematical problem datasets such as the MATH dataset (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021). This will establish a baseline performance level.

2. **Chain-of-Thought (CoT) Prompting**: I will implement CoT prompting (Wei et al., 2022), which encourages the model to generate intermediate reasoning steps before arriving at a final answer. This technique has been shown to improve reasoning capabilities in LLMs.

3. **Retrieval-Augmented Generation (RAG)**: I will incorporate RAG (Lewis et al., 2020) by providing the LLM with relevant context from mathematical textbooks, lecture notes, and other resources. This context will be retrieved based on the problem at hand, potentially improving the model's understanding and solution accuracy.

4. **Implementation and Integration**: Building upon the Python code I developed during the Kaggle competition, which utilized self-consistency, code execution, and advanced prompting techniques, I will integrate RAG

and CoT into a unified framework. This will involve developing modules for context retrieval, prompt engineering, and reasoning evaluation. The previous code focused on generating Python code to solve mathematical problems and executing it to obtain the final answer. By enhancing this framework with RAG and CoT, I aim to improve both the reasoning process and the accuracy of solutions.

5. **Evaluation and Analysis**: I will evaluate the enhanced model's performance on college-level mathematical problems, analyzing both the accuracy of the final answers and the correctness of the reasoning steps. I will compare results with the baseline and previous work to assess improvements.

My approach differs from previous work by combining RAG and CoT techniques specifically for mathematical problem-solving at the college level, and by providing additional domain-specific context to the LLM. I aim to investigate whether these methods can synergistically enhance the LLM's reasoning capabilities.

Topics from NLP Course Covered

- **Large Language Models (LLMs)**: Understanding and utilizing LLM architectures.
- **Prompt Engineering**: Crafting effective prompts to elicit desired responses from LLMs.
- **Retrieval-Augmented Generation (RAG)**: Combining retrieval mechanisms with generation models.
- **Chain-of-Thought (CoT) Reasoning**: Encouraging step-by-step reasoning in LLMs.

What baseline algorithms will you use? My baseline will consist of evaluating the LLMs using standard prompting without any additional context or special techniques. Specifically:

- **Zero-Shot Prompting**: Directly asking the LLM to solve the problem without any intermediate steps or additional context.
- **Standard Chain-of-Thought (CoT) Prompting**: Using CoT prompting without retrieval augmentation to see the effect of reasoning prompts alone.
- **Baseline LLM Performance**: Measuring the performance of the LLMs on the datasets without any modifications, establishing a minimal performance benchmark.

These baselines will help assess the effectiveness of my proposed methods compared to simple and straightforward approaches.

2.1 Course of Action

I plan to complete the project with the following steps:

1. *Data Acquisition and Pre-processing*
   - Collect datasets of college-level mathematical problems from HuggingFace and other sources.
   - Pre-process the data to ensure it is suitable for input to the LLMs.
2. *Baseline Evaluation*
   - Implement the baseline algorithms and evaluate the LLMs' performance.
   - Analyze results to identify key challenges and areas for improvement.
3. *Implementation of CoT and RAG*
   - Develop modules for Chain-of-Thought prompting.
   - Set up the retrieval system for Retrieval-Augmented Generation using textbooks and lecture notes.
   - Integrate CoT and RAG into the existing codebase from the Kaggle competition.
4. *Model Training and Fine-tuning*
   - Fine-tune the LLMs with the new framework.
   - Perform experiments to optimize parameters.
5. *Evaluation and Error Analysis*
   - Evaluate the enhanced model on the test datasets.
   - Conduct in-depth error analysis to understand failures and limitations.
6. *Final Report Writing*
   - Document methodology, experiments, results, and analysis.
   - Prepare the final report in LaTeX format.

3 Data and Compute

Data Sources

- **Mathematical Problem Datasets**: I will use publicly available datasets such as the MATH dataset (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), and other similar datasets hosted on HuggingFace.
- **Textbooks and Lecture Notes**: For retrieval augmentation, I will use digital versions of college-level mathematics textbooks and lecture notes, either from open-source materials or with appropriate permissions.

Data Availability

- All datasets are freely available and can be easily accessed and downloaded.
- Textbooks and lecture notes will be sourced ensuring compliance with copyright laws.

Compute Resources

- **Compute Requirements**: Running LLMs, especially with retrieval and reasoning components, can be computationally intensive.
- **Available Resources**: I plan to use Kaggle's 30+20 hours of TPU and GPU compute, as well as any available university resources (UCSD DataHub) or cloud credits.
- **Feasibility**: The compute resources are sufficient for running experiments at the required scale. I will optimize my code to ensure efficient use of compute.

References

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*.

Hendrycks, D., Burns, C., Kadavath, S., Arora, S., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems*.

Investments, X. (2024). Ai mathematical olympiad - progress prize 1. https://kaggle.com/competitions/ai-mathematical-olympiad-prize. Kaggle Competition.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., and Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.