10623 Project Proposal
Improving Math Problem Solving with Long Context LLMs

Anish Kiran Kulkarni (anishkik@andrew.cmu.edu)
Rajeev Veeraraghavan (rveerara@andrew.cmu.edu)
Ajay Mittur (amittur@andrew.cmu.edu)

1 Introduction

As part of this project, we would like to explore the potential improvements to performance of LLMs generated due to increasing context size of most LLMs. Specifically we would like to explore the improvements for in context learning that can be achieved by incorporating more information in the prompt made possible due to the larger context size. Inspired by some recent research **?**, we would like to explore the benefits of many shot learning on solving math problems. We would like to verify the improved accuracies demonstrated in **?** for open source LLMs in solving math problems.

2 Dataset & Task

We will use the MATH **?** and GSM8K **?** datasets to evaluate whether many-shot prompting improves performance over few-shot prompting. We measure accuracy with exact match of LLM generated answers to ground truth solutions.

2.1 MATH

The MATH dataset **?** contains 12,500 high school-level math problems from competitive math events. Each problem requires machine learning models to produce a final answer, such as $\frac{2}{3}$, in a normalized form. This normalization makes answers unique, allowing exact match metrics instead of heuristic metrics. Problems in the dataset are categorized by difficulty on a scale of 1 to 5 and span seven mathematical domains like number theory, algebra, and probability. The main goal of the dataset is to evaluate a model's ability to tackle complex mathematical problems and provide concise solutions.

2.2 GSM8K

GSM8K **?** consists of 8500 high quality grade school math problems created by human problem writers. The dataset is split into 7500 training problems and 1000 test problems. Each problem takes between 2 and 8 steps to solve. The human written solutions are in natural sentence form and perform a sequence of basic math operations in each step. The final step contains the answer. Unlike the MATH dataset, the problems are not classified by topics. However, the answers are exact. An exact match accuracy metric is employed to evaluate performance on this dataset as well.

3 Related Work

3.1 Using the long context for many shot in context learning

In many shot in context learning **?** the authors analyze the benefit of the long context available in Gemini 1.5 Pro LLM. They explore two settings - reinforced in context learning and unsupervised learning. In reinforced in context learning, the authors model generated chain of thoughts in examples provided in the prompt and in unsupervised in context learning they remove reasoning and only include domain specific examples in the prompt. The authors observe and report improvement in performance for various tasks using many shot in context learning. Improvements in accuracy on GSM8K **?** are from 84% to 93% and MATH **?** from 50% to 58.1%.

3.2 Analyzing the benefit of longer context overall

In In-Context Learning with Long-Context Models: An In-Depth Exploration **?** the authors analyze in context learning for LLMs with larger context size. They identify that performance continues to increase with a large number of examples in the prompt. They identify that example retrieval provides diminishing returns as the length of the context increases. They also find long context in context learning less sensitive to the order of examples as compared to short context. They conclude that overall understanding of in context learning especially in view of the increasing context size remains incomplete and requires further work for more hypothesis validation.

3.3 Analyzing RAG performance using long context

In Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG **?** the authors evaluate the impact and benefits of long context on performance of retrieval augmented generation. The authors observe that having more data in the prompt improves quality initially but it reduces as the as the retrieval size keeps increasing. The authors observe that the irrelevant passage negatively affects retrieval augmented generation performance and proppose optimization approaches with and without training to improve performance.

3.4 Analyzing impact of information location in input for long context

In Lost in the Middle: How Language Models Use Long Contexts **?** the authors try to analyze how LLMs tend to use the longer context typically availble with recent LLMs. The authors analyze performance using two tasks - multi document question answering and key value retrieval. They observe that the performance on the task is very dependent on the location of the relevant information. The observe that performance is highert when relevant information is at the beginning of the input and performance reduces as the relevant information is further towards the middle of the input. The performance again improves as the relevant information is towards the end of the input.

4 Approach

We aim to re-implement from scratch the work presented in **?** - where supervised and unsupervised many-shot prompting improved performance on MATH and GSM8K datasets in Gemini LLM models. The many-shot prompts' tokens' lengths were of the order of 100k. This means that we should be able to replicate their results on models with maximum context window lengths of 128k tokens.

We propose to study the effect of many-shot prompts on open source models like Llama. We will evaluate whether unsupervised many-shot prompting - where the prompt includes up to 500 questions, but no answers - outperforms regular few-shot prompting with both questions and answers. We will also evaluate whether synthetically generated answers improve accuracy, irrespective of their quality.

The results that we plan to replicate are the consistent improvements over few-shot prompting achieved by both supervised (re-inforced) and unsupervised many-shot prompting (for an average of a 7.9% improvement) on MATH. Even if our average improvement is not the same, we expect to see a similar trend where many-shot prompting (unsupervised and re-inforced) beats regular in-context-learning over a range of a number of many-shot prompts. We also expect to replicate the finding that the many-shot prompts obtained from MATH transfer to GSM8K and improve performance slightly even if GSM8K's question distribution is different.

For efficiency, we will use existing frameworks (such as DsPY **?**) to obtain structured outputs and implement prefix prompt caching to save computation and or cost when adding many-shot examples. To do prefix prompt caching, we will only add new examples (shots) as we increase the number of shots in our many-shot ablations.

5 Expected outcomes

We hypothesize, that with the increase in context lengths of LLMs, we would be able to improve an LLMs reasoning capabilities by simply scaling the number of in-context examples, in the order of hundreds. Further, we also take a step further and posit that the quality of these in-context examples is not important and synthetically generated examples will realize similar performance gains as human annotated data. Finally, we also want to analyze and understand if there are a smaller set of many-shot examples that provide most of the performance gains.

To validate this hypothesis, the experiments we plan to run are:

- Measure model performance differs with different numbers of many shot examples.
- Compare model performance using human annotated few shot prompts and our many shot prompts created with synthetic data.
- Use synthetic data without any data filtering and validation, synthetic data without reference answers (unsupervised ICL), synthetic data with CoT reasoning steps.
- Evaluate different long context open source small and large LMs and measure their performances. We are particularly interested in leading models, Phi, Mistral, Llama 3.1

At the end of our experiments we hope to get results that support our hypothesis. That is, an improvement over existing benchmarks that use few-shot prompts on the datasets selected. Moreover, we also hope to validate our notion that data quality becomes less important as the number of in-context examples increases.

6 Plan

We will be collaborating closely for most part of the project through active working sessions and pair programming. However, we also realize the important of splitting up tasks and working in parallel at times. Hence, we have categorized the effort into three main areas: data generation, inference pipeline, evaluation and experimentation, and each of us will take up tasks spanning these areas respectively. As everyone in the team has extensive experience in natural language processing, reinforcement learning, and data generation, we do not plan on isolating tasks that a single individual picks up. That way everyone gets to work on every component of the project as well. We plan to check in every 3 days to discuss our progress, blockers or any issues anyone is facing. By the midway executive summary, we plan to have completed the re-implementation of the paper, so that we can focus on improvement and experimentation for the remainder of the project. These can also be interpreted as the major milestones for our project. That is, 1) pipeline re-implementation using open LLMs 2) experimentation and improvement using our proposed methods. Because there are only roughly 4 weeks to complete the project, we felt two milestones, roughly 2 weeks apart was the best way to proceed. We will break it down further into granular tasks once we start implementing the project.

References

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning, 2024. URL: `https://arxiv.org/abs/2404.11018`, `arXiv:2404.11018`.

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, 2021. URL: `https://arxiv.org/abs/2110.14168`, `arXiv:2110.14168`.

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.

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.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. In-context learning with long-context models: An in-depth exploration, 2024. URL: `https://arxiv.org/abs/2405.00200`, `arXiv:2405.00200`.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O. Arik. Long-context llms meet rag: Overcoming challenges for long inputs in rag, 2024. URL: `https://arxiv.org/abs/2410.05983`, `arXiv:2410.05983`.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023. URL: `https://arxiv.org/abs/2307.03172`, `arXiv:2307.03172`.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into self-improving pipelines, 2023. URL: `https://arxiv.org/abs/2310.03714`, `arXiv:2310.03714`.