10623 Midway Executive Summary
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

We explore how increasing context lengths in modern LLMs can improve mathematical reasoning through many-shot in-context learning. Recent work has shown that using hundreds of examples in prompts can significantly boost performance on math tasks. We aim to validate these findings using open-source LLMs and investigate whether synthetic examples can match human-annotated data. Our approach focuses on re-implementing and extending the many-shot prompting techniques from **?** using models like Llama with 128k context windows. We evaluate on the MATH and GSM8K datasets, expecting to demonstrate that increasing the number of in-context examples improves performance regardless of example quality. Our preliminary results with baseline few-shot prompting show promise, and we plan to systematically compare supervised and unsupervised many-shot approaches.

2 Dataset & Task

We evaluate on two mathematical reasoning datasets - MATH and GSM8K. Our primary metric is exact match accuracy between model outputs and ground truth solutions.

2.1 MATH

The MATH dataset **?** contains 12,500 high school-level math problems from competitive math events. Each problem requires producing a normalized final answer (e.g. $\frac{2}{3}$). Problems are categorized by difficulty (1-5) across seven mathematical domains. The dataset tests complex mathematical reasoning and step-by-step problem solving.

2.2 GSM8K

GSM8K **?** comprises 8,500 grade school math word problems (7,500 train, 1,000 test). Problems require 2-8 solution steps, with human-written natural language solutions. While problems aren't categorized by topic, answers are exact for reliable evaluation.

3 Related Work

3.1 Many-Shot In-Context Learning

Recent work **?** demonstrated significant gains using many-shot prompting with Gemini 1.5 Pro's long context. They explored both supervised and unsupervised approaches, showing 7.9-9

3.2 Long Context Benefits

Studies **?** show performance scales with more examples, though with diminishing returns. Long-context models are less sensitive to example ordering compared to short-context ones.

3.3 RAG with Long Context

Research **?** reveals initial benefits from increased context in retrieval-augmented generation, but performance plateaus with too many retrieved passages.

3.4 Information Positioning Effects

Analysis **?** shows performance varies based on information location - stronger when relevant content appears at prompt start/end versus middle.

4 Approach

We are implementing a systematic evaluation of many-shot prompting techniques using open-source LLMs with 128k context windows. Our key components include:

- Baseline: Few-shot prompting with 3-5 examples
- Unsupervised many-shot: Up to 500 questions without answers
- Supervised many-shot: Questions with synthetic answers
- Efficient implementation using DsPY framework and prefix prompt caching

5 Experiments

We will conduct the following experimental evaluations:

5.1 Baseline Results

Current few-shot prompting results on our test sets: [Results to be added]

5.2 Scaling Analysis

We will evaluate performance vs number of examples (10 to 500): [Table/plot to be added showing accuracy vs. number of examples]

5.3 Example Quality Impact

Comparison of different example types: [Table comparing human-annotated vs synthetic data performance]

5.4 Model Size Effects

Performance across different model scales: [Table comparing results across model sizes]

6 Plan

Remaining project timeline:

- Week 1-2 (All members):
  - Complete baseline implementation
  - Initial experiments with few-shot prompting
- Week 3 (Anish, Rajeev):
  - Implement many-shot variations
  - Synthetic data generation
- Week 4 (Ajay, Rajeev):

– Run full experimental suite

– Analysis and documentation

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.

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.

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.

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.