# Improving Math Problem Solving with Long Context LLMs


10-423/623 Generative AI Course Project
Ajay Mittur, Rajeev Veeraraghavan, Anish Kiran Kulkarni
(amittur, rveerara, anishkik)@andrew.cmu.edu


December 14, 2024

## 1 Introduction

Recent advancements in large language models (LLMs) have significantly increased their context sizes (Ding et al. [2024]), enabling more extensive in-context learning capabilities. With these larger context windows, we can now feed models many more examples at once through many-shot prompts, potentially helping them perform better on difficult tasks.

Research from Google DeepMind (Agarwal et al. [2024]) examines the impact of many-shot in-context learning on proprietary Gemini models (Team et al. [2024]). Their work highlights the advantages of many-shot in-context learning in solving mathematical problems by using up to 500 ICL shots.

They find a net **7.9%** improvement over baseline 4-shot prompting on both MATH Hendrycks et al. [2021] and GSM8K, with their best performing prompts using 125 and 500 shots respectively. The context lengths of these prompts were on the order of $10^5$ or 100,000 tokens. Inspired by their results, we examine whether their findings apply to open-source LLMs with smaller context sizes. Since open source LLMs with context window sizes of 128,000 tokens exist, we replicate their study using many-shot prompts of similar sizes.

By using the full context window size of Llama3.1-8B-Instruct (Dubey et al. [2024]), we explore if adding more examples into the prompt - with or without verified answers and/or explanations - can boost model accuracy on MATH (Hendrycks et al. [2021]). Our method includes comparing many-shot to traditional few-shot prompting. We also look into how synthetic data can enhance many-shot ICL performance. Through detailed experiments, we validate whether the performance improvements seen in proprietary Gemini models (Agarwal et al. [2024]) can crossover to an open source LLMs with a smaller context window size.

We evaluate performance on both datasets in three distinct settings - which we explain in further detail in our methodology section:

- **Supervised many-shot ICL** - The ICL examples (shots) include both questions and answers (with reasoning steps included) from the MATH Hendrycks et al. [2021] Train dataset (which is the training subset of MATH).
- **Unsupervised many-shot ICL** - The ICL shots include only questions (without any answers or explanations) from the MATH Train dataset.
- **Synthetic data augmented / Reinforced many-shot ICL** - The ICL shots include synthetically generated questions and answers (with reasoning steps) generated based on questions from the MATH Train dataset.

Early results show that many-shot prompting, particularly in unsupervised settings, greatly boosts problem-solving accuracy. We also looks at what this means for the future of LLMs and their applications.

## 2 Dataset & Task

We use the MATH Hendrycks et al. [2021] dataset. We measure accuracy, gauged through exact match between the answers generated by the LLM, and the ground truth solutions. Given that answers might be in latex, we parse these prior to conducting an exact match. Specifically, we utilize the evaluation function from the Minerva paper Lewkowycz et al. [2022], with the source code available here, closely aligning with Meta's evaluation for Llama 3.1. Hosting 12,500 high school-level math problems from competitive events, the MATH dataset Hendrycks et al. [2021] challenges models to generate final answers, like $\frac{2}{3}$, in a standardized form. This standardization ensures uniqueness of answers, facilitating the use of exact match metrics. The problems span seven domains, including number theory and algebra, and are graded on a difficulty scale from 1 to 5. The dataset's primary aim is to test a model's prowess in solving intricate mathematical problems and delivering succinct solutions.

### 2.1 MATH500

MATH500 (HuggingFace) is a subset of 500 problems from the full MATH Hendrycks et al. [2021] Test subset. They are selected to be a representative sample of the full test set, and are often used as standalone benchmark with researchers reporting performance on it. Google DeepMind's team also reported on the MATH500 dataset. We do the same for consistency. MATH500 was used for evaluation only. All ICL examples were either selected or synthetically generated from the MATH Train set. Since MATH500 and MATH Train are disjointed, we avoid any contamination and data leakage.

Additionally, we split the MATH500 dataset into a smaller subset featuring solely numeric answers—either integers or floats. This split serves two purposes: it allows for evaluation without the need for latex parsing or reliance on non-exact match metrics, and it can enable us to test if few and many-shot prompting with this "numerical answer" subset improves performance on out-of-distribution GSM8KCobbe et al. [2021] problems, compared to using a broader selection of examples. This method facilitates a comparative evaluation of performance across different data distributions on a more granular level.

# 3 Related Work

## 3.1 Increasing context size of LLMs

Long context in LLMs have been getting a lot of attention from research in recent times and multiple approaches have been tried to increase the capacity of LLMs to support longer contexts. LongRope Ding et al. [2024] explores increasing the context length to more than 2 million by using a step wise approach to positional interpolation. PoSE Zhu et al. [2024] tries to extend the context window without having to undertake full finetuning. Another approach Dong et al. [2024] tries to extend the context window using positional vector replacement.

## 3.2 Effective context size

The long context also comes with challenges. Recent research An et al. [2024] identifies that the effective context length of an LLM falls short, often to half of the original value. The authors propose a new embedding technique which they call String embeddings to increase performance on long context benchmark tasks.

## 3.3 Using the long context for many shot ICL

In many shot in context learning Agarwal et al. [2024] 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 Cobbe et al. [2021] are from 84% to 93% and MATH Hendrycks et al. [2021] from 50% to 58.1%.

## 3.4 Analyzing the benefit of longer context

In In-Context Learning with Long-Context Models: An In-Depth Exploration Bertsch et al. [2024] 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.

## 3.5 Analyzing RAG using long context (1)

In Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG Jin et al. [2024] 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.

## 3.6 Analyzing RAG using long context (2)

In Long Context RAG Performance of Large Language Models Leng et al. [2024], the authors identify the effect of longer context on 20 major LLMs. They observe that retrieving more documents can improve performance but the improvement is generally lost as the context increases above 64K tokens.

## 3.7 Analyzing impact of information location in input for long context

In Lost in the Middle: How Language Models Use Long Contexts Liu et al. [2023] 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. They observe that performance is highest 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.

### 3.8 Does long context help ICL

In Long-context LLMs Struggle with Long In-context Learning Li et al. [2024], the authors introduce a new benchmark for in context learning called Long-ICLBench to analyze the benefit of long context to in context learning for LLMs. They observe that the longer context helps with easy tasks but on difficult tasks almost all LLMs fail. They also observe that the LLMs tend to be more attentive to examples towards the end of the prompt.

### 3.9 Is long context necessary?

In Are Long-LLMs A Necessity For Long-Context Tasks? Qian et al. [2024], the authors provide a framework called LC Boost for using short context LLMs to solve long context problems. Based on the results observed with small context models using this framework, the authors argue that long context is not necessary and short context LLMs can also solve a lot of long context problems using this framework.

### 3.10 Does ICL help with instruction following?

In Is in-context learning sufficient for instruction following in LLMs? Zhao et al. [2024], the authors try to compare the performance of in context learning with instruction finetuning and try to analyze if in context learning can help an LLM in alignment. The authors provide an insight that providing high quality examples in the context can help the model performance increase towards that of an instruction fine tuned model.

## 4 Methods

We re-implement from scratch the work presented in Agarwal et al. [2024]. Instead of Gemini, we evaluate the effect of many shot prompting on open source Llama3.1-8B-Instruct Dubey et al. [2024]. The three methods we evaluate are supervised, unsupervised and re-inforced ICL with increasing number of shots.

### 4.1 Baseline Approach

We re-evaluated baselines on our own and compared if they match reported results. There are two baselines to consider, since we measure improvement over both zero/few-shot ICL and supervised ICL, in general. Hence, one baseline is Llama-3.1-8B-Instruct's accuracy with zero, 3 and 5 shot ICL (with questions and reasoning steps included in the ICL examples). The second baseline is supervised ICL, regardless of number of shots. Since Agarwal et al. [2024] claim that unsupervised and re-inforced ICL beat supervised ICL once we include enough shots, our baseline includes many-shot supervised ICL.

### 4.2 Main Methods

First, we evaluate unsupervised many-shot prompting - where the ICL examples include up to 250 questions from MATH Train dataset, without including any answer. While Agarwal et al. [2024] use 4, 10, 25, 50, 125, 250, 500 shots, we also include 75, 100 shot experiments to interpolate the results better, and estimate where peak performance occurs. We do not evaluate 500-shot ICL since accuracy shows a clear decreasing trend starting from 125 shots. We instead evaluate to a maximum of 250 shots.

Then, we evaluate the same for re-inforced ICL. Here, the ICL shots include questions and answers with reasoning. Both questions and answers are generated synthetically by Llama-3.1-70B Dubey et al. [2024] using examples from the MATH Train dataset. During inference, the ICL shots are sampled at random, similar to how they are sampled for the other methods.

## 5 Experiments

As described in the dataset section, we report performance on a "numeric-answer" subset of the MATH500 dataset (HuggingFace), consisting of 323 questions. All our experiments were run on Llama3.1-8B-Instruct Dubey et al. [2024] model using an NVIDIA A100 Tensor Core GPU with 80GB RAM 2000 GB/s bandwidth. We downloaded the model from HuggingFace hub and ran all inference locally using VLLM Kwon et al. [2023]. VLLM implements prefix prompt caching to save computation and cost when adding many-shot examples. To ensure efficient prefix prompt caching, we appended new example shots to our prompt and re-used the ICL shots across evaluation questions to save computation. Unless stated otherwise in our results table, these were the experimental settings used (per row entry).

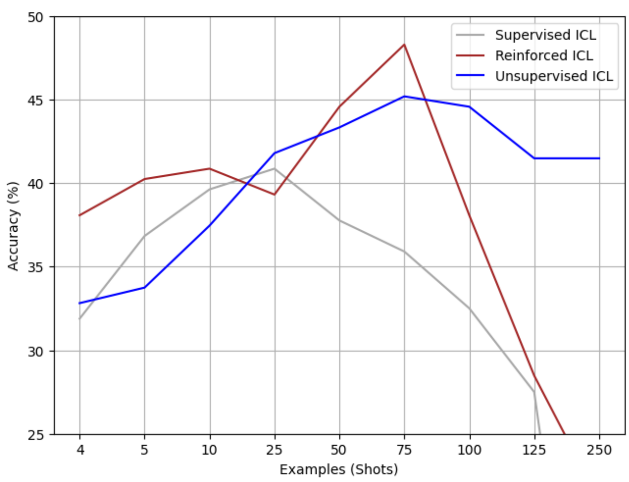


Figure 1: Accuracy trend by amount of examples

Results of our experiments are presented in table

| Model | Prompt | Dataset | Subset | Accuracy (%) |
|---|---|---|---|---|
| Llama 3.1 8B Instruct Dubey et al. [2024] | Zero Shot CoT | MATH | All | 51.9% |
| Llama 3.1 8B Instruct (Our Eval) | Zero Shot CoT | MATH | All | 47.6% |
| Llama 3.1 8B Instruct (Our Eval) | Zero Shot CoT | MATH | Numeric | 44.46% |
| Gemini Ultra Agarwal et al. [2024] | 250 Shot CoT | MATH | MATH500 | 58.8% |

Table 1: Baseline results

| Prompt | Source | Type | Dataset | Subset | Accuracy (%) |
|---|---|---|---|---|---|
| 3 Shot CoT | Synthetic | Supervised | MATH | All | 31.8% |
| 5 Shot CoT | MATH Test | Unsupervised | MATH | Numeric | 35.19% |

Table 2: Initial experiment resutls with Llama 3.1 8B Instruct Dubey et al. [2024]

| Examples | Unsupervised Acc | Supervised Acc | Reinforced Acc | Approximate Runtimes (minutes) |
|---|---|---|---|---|
| 4 | 32.82 | 31.89 | **38.08** | (2, 5, 5) |
| 5 | 33.75 | 36.84 | **40.25** | (3, 6, 6) |
| 10 | 37.46 | 39.63 | **40.87** | (6, 11, 11) |
| 25 | **41.80** | 40.87 | 39.32 | (10, 20, 20) |
| 50 | 43.34 | 37.77 | **48.30** | (20, 38, 36) |
| 75 | **45.20** | 35.91 | 38.08 | (30, 60, 55) |
| 100 | **44.58** | 32.51 | 38.08 | (42, 95, 82) |
| 125 | **41.49** | 27.50 | 28.38 | (50, 120, 105) |
| 250 | **41.49** | - | 21.00 | (90, -, 180) |

Table 3: Experiment results (Accuracy %) with Llama 3.1 8B Instruct Dubey et al. [2024] on MATH500 HuggingFace numeric subset with CoT prompting

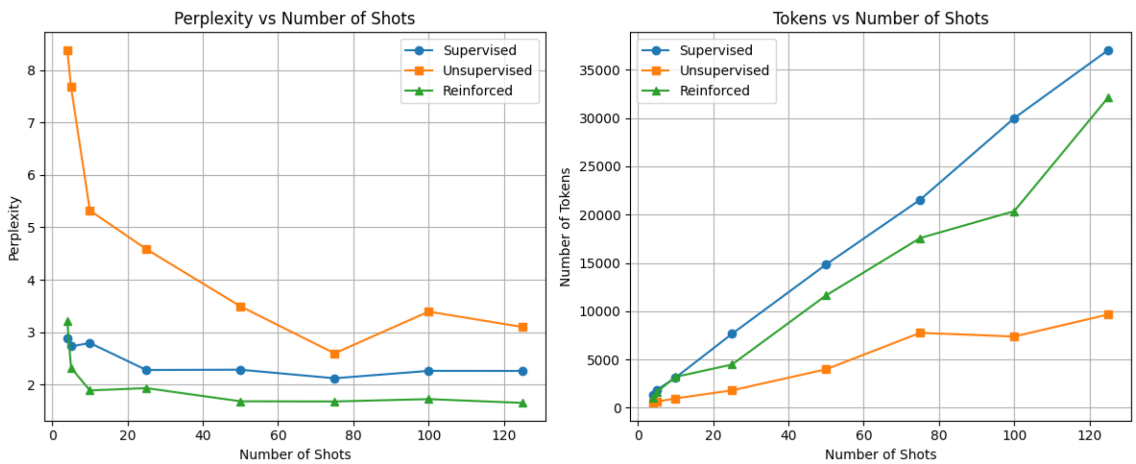


Figure 2: Trend in perplexity and prompt size as the amount of examples changes

3 and a plot of the results trends is available in figure 1. We observe that as the number of examples (shots) provided to the model increases, unsupervised ICL outperforms supervised ICL as well as reinforced ICL. Further observation and analysis of these results is presented in the theorotical foundations part of the research log section.

## 6 Code Overview

We have included screenshots in the appendix section. Our code is available at https://github.com/Rajeevveera24/manyshot-math and the gradescope code submission url is https://www.gradescope.com/courses/820031/assignments/5454735/submissions/297439187

**Dataset curation code** - figure 3 - extracted answers from text for MATH and MATH500 datasets and stored both questions, explanations and Ids locally in JSON files.

**Baseline code** - Figure 4 - Represents the initial setup to evaluate the model's performance on zero and few shot baselines.

**Experiment Logger (a and b)** - Figures 5 and 6 - Depict the mechanisms for tracking experiments and storing JSON results to ensure we can re-start crashed runs, and later analyze results.

**Synthetic data generation** - Figure 11 - Generates the synthetic questions and answers needed for re-inforced ICL.

**Unsupervised ICL code** - Figure 7 - Implements unsupervised in-context learning by retrieving only questions from the MATH Train (code for running re-inforced ICL and supervised ICL is similar).

**Perplexity Calculation** - Figure 9 - Highlights the code used to compute perplexity of the input prompts.

**Result Plotting code** - Figure 10 - Visualizes the results of the experiments for analysis and interpretation.

# 7 Timeline

Our timeline is presented in table 4.

# 8 Research Log

- Evaluation Dataset Selection - We were initially reporting accuracy on the entire MATH Test dataset with 3000+ examples. Later, we realized that Agarwal et al. [2024] had evaluated on MATH500 instead and switched to using that instead of the full MATH Test. While this meant that we had to re-run our baselines again, it sped up experimentation since we were evaluating on lesser questions.

- Dataset curation - The dataset's examples contained final answers as part of a string that also had explanations. We had to parse the final answers for each example using regex matching. Further, we chose to focus on a numeric subset rather than rely on parsing Latex outputs from the model. "Numeric" subsets of MATH Train and MATH500 were derived from the full sets where the answers to questions in these subsets were purely numeric. We ran some of our initial experiments on the full subsets though, before changing over to our numeric subsets later.

- Experiment logging - We wrote a common experiment module capable of writing results to a JSON file periodically while an experiment was in progress. This allowed us to restart runs from the last saved checkpoint. This was crucial since our run-times often crashed when we ran experiments with 100+ ICL examples.

- Using VLLM in favor of DsPY - While we initially favored using DsPYKhattab et al. [2023] for its ease of chaining model outputs and type parsing, we eventually switched to using VLLM directly and wrote the code to parse model outputs ourselves. VLLM was significantly faster, allowing us to batch our inputs and iterate faster. However, we could not implement KV caching which is a key component for manyshot ICL to work in a practical setting.

- Synthetic data generation - We approached this problem with the expectation that open source models would've been as good as large closed source models in synthetic data generation with the number of studies that had been conducted. However, upon starting to actually implement simple MATH generation, we quickly realized that to generate good representative date, there would be a lot of moving components involved, such as a verifier, solver, etc. In short, synthetic math data generation was another re-search direction in itself.

- Theoretical Foundations - Once we replicated Google DeepMind's (Agarwal et al. [2024]) results, we realized that the base paper too has very weak theoretical justifications for why it's findings. Further examining the literature on long-context ICL, we measured two metrics that point to why re-inforced (synthetic) and unsupervised (question-only) ICL works. One is that the perplexities of the input prompts (fig. 2) - which contain the ICL examples - are lower for the synthetic inputs than for the supervised input prompts. This effect increased with the number of shots and possibly explains why task performance was better. The second metric is the percentage of the context window size of 128,000 tokens used by the input prompts. The unsupervised input prompts were significantly shorter than both the synthetic and supervised prompts (fig. 2). This meant that attention was spread over a smaller set of tokens allowing the model to sample the correct token at each decoding step. This is is line with findings that most LLMs perform optimally when input prompts are significantly shorter than maximal context window lenghts . We plan to run further experiments to verify if both hypotheses hold true in more diverse settings, at least empirically.

- Out of data distribution generalization - We had initially planned on applying the techniques which worked on MATH to GSM8K as well. This would have been especially helpful to evaluate if the synthetic data generated from MATH transferred to improvement in GSM8K, similar to Agarwal et al.

| Activity | Start | End | Hours |
|---|---|---|---|
| **Literature Review** | | | |
| Reviewing Papers for Ideas | Nov 5 | Nov 12 | 12 |
| Reading baseline paper in depth | Nov 11 | Nov 15 | 6 |
| Reading related papers to baseline | Nov 13 | Dec 5 | 16 |
| **Dataset and set up** | | | |
| Dataset Curation | Nov 13 | Nov 25 | 8 |
| Environment and dependency setup | Nov 13 | Nov 25 | 4 |
| **Implementation** | | | |
| Writing experiment code | Nov 20 | Dec 11 | 11 |
| Reading eval, sympy, dspy and vllm code | Nov 20 | Dec 3 | 9 |
| Synthetic data generation pipeline | Nov 20 | Nov 30 | 15 |
| **Experimentation** | | | |
| Baseline and initial experiments | Nov 18 | Nov 25 | 11 |
| Running main experiments for all 3 | Dec 3 | Dec 12 | 20 |
| Analysis, perplexity calculation and visualization | Dec 6 | Dec 10 | 5 |
| **Documentation** | | | |
| Writing project proposal | Nov 13 | Nov 15 | 4 |
| Writing Midway Executive Summary | Nov 22 | Nov 25 | 7 |
| Creating & Presenting Final Poster | Dec 5 | Dec 10 | 18 |
| Writing Final Executive Summary | Dec 7 | Dec 13 | 12 |
| **Total Hours** | | | 158 |

Table 4: Timeline and effort breakdown for project activities

[2024]'s findings. We could not experiment with this due to the limited time that we had.

# 9 Conclusion

Our findings with many-shot ICL on Llama-3.1-8B-Instruct align with the conclusions presented by Google DeepMind Agarwal et al. [2024] on Google' Gemini family of models. Specifically, we observe that many-shot ICL enhances performance on MATH when using unsupervised and reinforced over traditional few-shot ICL. As the number of shots crosses 25, the latter is surpassed by both. Our results might be explained by the lower perplexity of re-inforced ICL and the lower context window utilization of unsupervised ICL. In both unsupervised and re-inforced ICL, there are diminishing returns after 125-shot ICL with accuracy decreasing beyond that mark, which is in line with the findings of An et al. [2024].

## 9.1 Potential Applications

Unsupervised and synthetic ICL, when extendable to other tasks, can reduce the bottleneck of needing human annotated data in a prompt. Thus, since performance on any end task can be improved using the same long ICL prompt, inputs can be cached and attention weights pre-computed using prefix caching. The end task or question can be appended at the end of the prompt while inferencing. This has many potential appications in data scarce domains.

## 9.2 Future Work

- Evaluating performance to GSM8K - To evaluate if the synthetic data generated here can help out of distribution performance on a similar math problem solving dataset.
- Finding where the true input length to context window length bottleneck lies - By including the test question and its answer in our many-shot supervised examples, and increasing the number of shots, we can get an idea of the input length at which an LLM fails to "retrieve" the correct answer (since the answer is already present in the prompt). This is a truer measure of the impact of increasing input lenghts on performance.
- Category and Level Wise Ablations - We can further verify if selecting ICL examples based on the true category or level of the test question has any impact on accuracy. Our bias would be that selecting questions and answers from the same level and category would most help ICL accuracy.

In conclusion, our findings align with Agarwal et al. [2024]'s that unsupervised and re-inforced ICL counter-intuitively outperform supervised ICL, given enough examples. Further experimentation could give more empirical evidence to support our hypotheses on why it happens. We believe our study contributes to ongoing research on longer contexts in LLMs.

## 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.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. Why does the effective context length of llms fall short?, 2024. URL https://arxiv.org/abs/2410.18745.

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.

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.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

Zican Dong, Junyi Li, Xin Men, Wayne Xin Zhao, Bingbing Wang, Zhen Tian, Weipeng Chen, and Ji-Rong Wen. Exploring context window of large language models via decomposed positional vectors, 2024. URL https://arxiv.org/abs/2405.18009.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yi-wen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

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.

HuggingFace. Math500. URL `https:`

//huggingface.co/datasets/qq8933/MATH500.

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.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Quinn Leng, Jacob Portes, Sam Havens, Matei Zaharia, and Michael Carbin. Long context rag performance of large language models, 2024. URL https://arxiv.org/abs/2411.03538.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning, 2024. URL https://arxiv.org/abs/2404.02060.

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.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Yujia Zhou, Xu Chen, and Zhicheng Dou. Are long-llms a necessity for long-context tasks?, 2024. URL https://arxiv.org/abs/2405.15318.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang,

Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kepa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Mom-

chev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Is in-context learning sufficient for instruction following in llms?, 2024. URL https://arxiv.org/abs/2405.19874.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Pose: Efficient context window extension of llms via positional skip-wise training, 2024. URL https://arxiv.org/abs/2309.10400.

# A Appendix

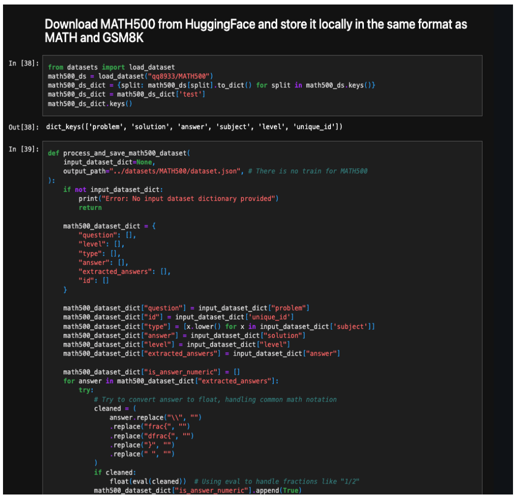


Figure 3: Dataset Curation

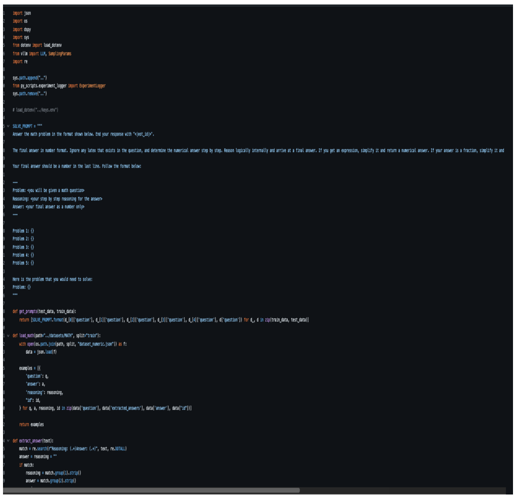

Figure 4: Baseline


Figure 5: Experiment Logger (a)


Figure 6: Experiment Logger (b)

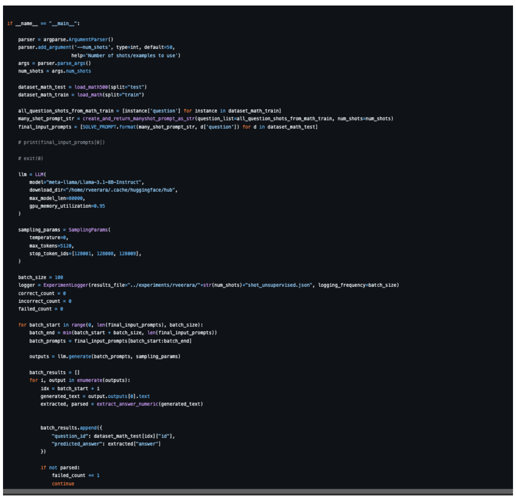

Figure 7: Unsupervised ICL

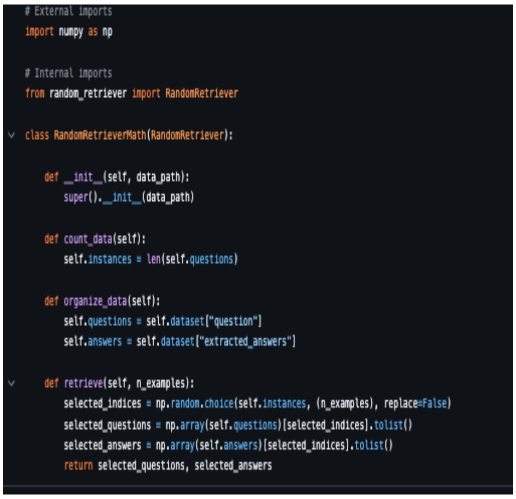


Figure 8: Question Retrievers


Figure 9: Calculate Perplexity

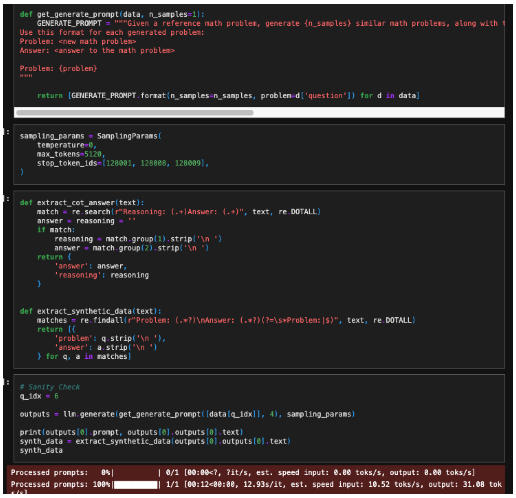


Figure 11: Plotting our results

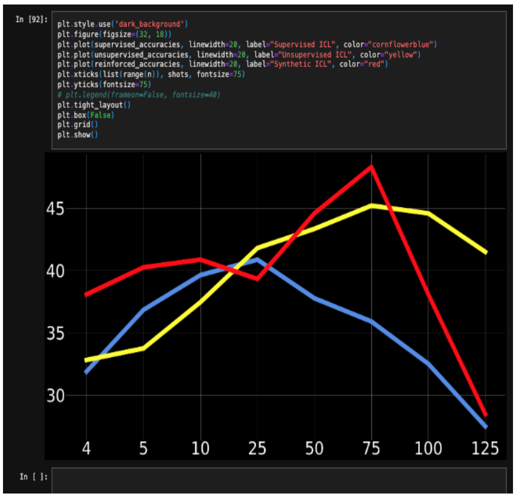


Figure 10: Plotting our results

## B Other results

| Model | Prompt | Source | Type | Dataset | Subset | Accuracy (%) |
|---|---|---|---|---|---|---|
| Llama 3.1 8B Instruct | 3 Shot CoT | Synthetic | Supervised | MATH | Test | 31.2 |
| Llama 3.1 8B Instruct | 5 Shot CoT | Synthetic | Supervised | MATH | Test | 34.6 |
| Llama 3.1 8B Instruct | 10 Shot CoT | Synthetic | Supervised | MATH | Test | 33.2 |
| Llama 3.1 8B Instruct | 50 Shot CoT | Synthetic | Supervised | MATH | Test | 30.3 |
| Llama 3.1 8B Instruct | 3 Shot CoT | Synthetic | Unsupervised | MATH | Test | 32.1 |
| Llama 3.1 8B Instruct | 5 Shot CoT | Synthetic | Unsupervised | MATH | Test | 33.9 |
| Llama 3.1 8B Instruct | 10 Shot CoT | Synthetic | Unsupervised | MATH | Test | 34.1 |
| Llama 3.1 8B Instruct | 50 Shot CoT | Synthetic | Unsupervised | MATH | Test | 32.3 |

Table 5: Many shot experiments with synthetic filtered and unfiltered data