A Careful Examination of Large Language Model Performance on Grade School Arithmetic

Hugh Zhang* **Jeff Da** **Dean Lee** **Vaughn Robinson** **Catherine Wu** **Will Song**

Tiffany Zhao **Pranav Raja** **Charlotte Zhuang** **Dylan Slack** **Qin Lyu**

Sean Hendryx **Russell Kaplan** **Michele (Mike) Lunati†** **Summer Yue†**

Scale AI

Abstract

Large language models (LLMs) have achieved impressive success on many benchmarks for mathematical reasoning. However, there is growing concern that some of this performance actually reflects dataset contamination, where data closely resembling benchmark questions leaks into the training data, instead of true reasoning ability. To investigate this claim rigorously, we commission *Grade School Math 1000* (GSM1k). GSM1k is designed to mirror the style and complexity of the established GSM8k benchmark, the gold standard for measuring elementary mathematical reasoning. We ensure that the two benchmarks are comparable across important metrics such as human solve rates, number of steps in solution, answer magnitude, and more. When evaluating leading open- and closed-source LLMs on GSM1k, we observe accuracy drops of up to 8%, with several families of models showing evidence of systematic overfitting across almost all model sizes. Further analysis suggests a positive relationship (Spearman's $r^2 = 0.36$) between a model's probability of generating an example from GSM8k and its performance gap between GSM8k and GSM1k, suggesting that some models may have partially memorized GSM8k. Nevertheless, many models, especially those on the frontier, show minimal signs of overfitting, and all models broadly demonstrate generalization to novel math problems guaranteed to not be in their training data.

1 Introduction

Improving reasoning in large language models (LLMs) is one of the most important directions of current research. As such, proper benchmarking of current LLM abilities is paramount for ensuring progress continues in the correct direction. Currently, the field typically relies on public benchmarks such as GSM8k (Cobbe et al. [2021]), MATH (Hendrycks et al. [2021b]), MBPP (Austin et al. [2021]), HumanEval (Chen et al. [2021]), SWEBench (Jimenez et al. [2024])). However, because LLMs are trained on large corpora of data scraped from the Internet, there are major concerns

*Correspondence to hugh.zhang@scale.com †equal senior authorship

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

https://doi.org/10.52202/079017-1485

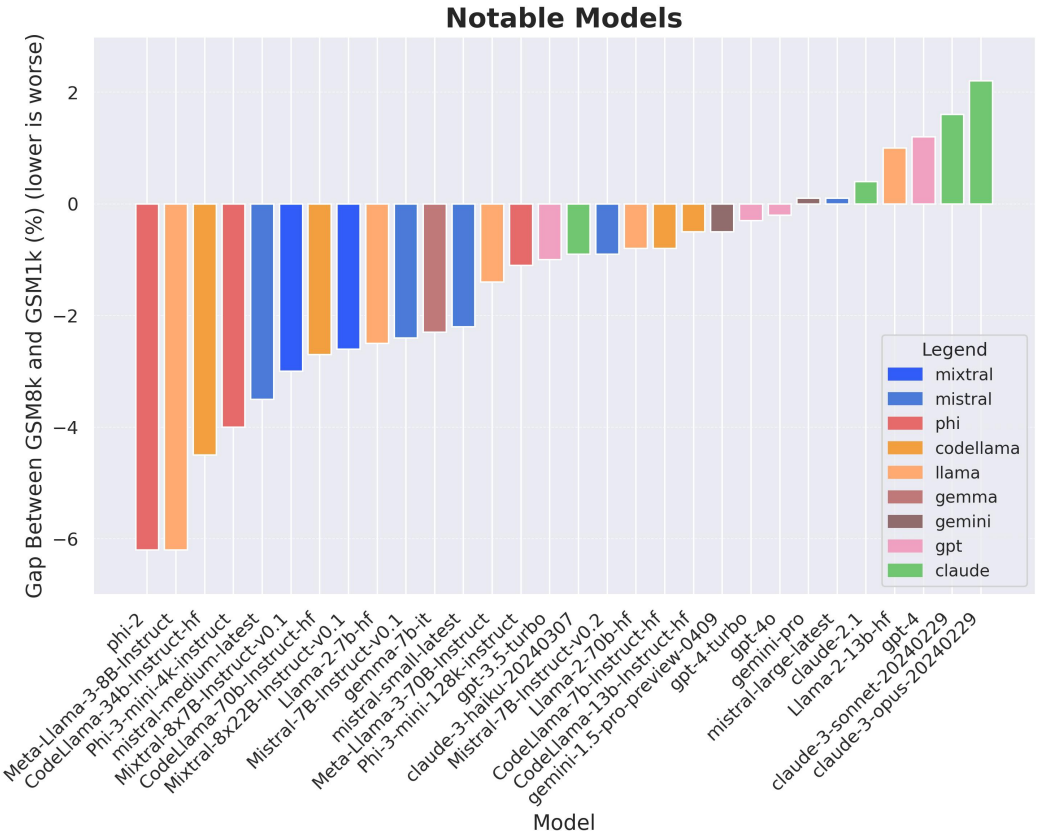

Figure 1: Notable models arranged by their drop in performance between GSM8k and GSM1k (lower is worse). We notice that Phi, Mistral and some models in the Llama family seem to be overfitting GSM8k, while models such as Gemini, GPT, and Claude show little to no signs of overfitting.

that such benchmarks may inadvertently include examples that closely resemble the questions found in such benchmarks. This contamination may result in models having weaker reasoning capabilities than otherwise believed, due to simply being able to repeat the correct answer that it previously encountered during pre- or post- training. To properly investigate the reasoning abilities of models, we commission GSM1k, a newly constructed collection of 1205 grade school level math problems designed to mirror that of GSM8k. We take extensive efforts to ensure that GSM1k has a similar distribution of difficulty to GSM8k to ensure an apples-to-apples comparison. These efforts are described in Section 3, alongside a detailed description of the data creation process. To mitigate worries about data contamination, we created GSM1k solely with human annotators, without assistance from any LLM or other synthetic data source.

Dataset	**Example**
GSM8k	James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?
GSM1k (ours)	Lee bought 6 shares of Delta stock at $40 per share. If he wants to make $24 from this trade, how much should Delta stock be per share when he sells?

Figure 2: Example from both the GSM8k dataset and the new GSM1k dataset (ours). We also provide an additional 50 examples from GSM1k in Appendix E.

https://doi.org/10.52202/079017-1485

We benchmark leading open- and closed-source LLMs on GSM1k, including GPT-4 (OpenAI et al. [2024]), Gemini (Team et al. [2024]), Claude, Mistral (Jiang et al. [2024, 2023]), Llama (Touvron et al. [2023a,b]), Phi (Gunasekar et al. [2023], Abdin et al. [2024]) and many more. Our analysis confirms the widespread suspicion in the field that many models are contaminated by benchmark data, with the worst models performing 8% worse on GSM1k compared to GSM8k. Additionally, our results suggest that several families of models show consistent evidence of overfitting for nearly all model versions and sizes. Further analysis finds a positive relationship (Spearman's $r^2 = 0.36$) between a model's probability of generating examples from GSM8k and its performance gap between GSM8k and GSM1k, strongly suggesting that one important component of this overfitting is that models have partially memorized examples from GSM8k. Nevertheless, our results find that all frontier models show minimal signs of overfitting. Additionally, we also find that all models, including the most overfit ones, are still capable of successfully generalizing to new mathematical grade school problems, albeit occasionally at lower rates than their benchmark numbers would suggest.

We do not intend to release GSM1k publicly at this time to prevent a similar problem of data contamination occurring in the future. However, we plan to run recurring evaluations of all major open- and closed- source releases and to continually update our results. We will also open source our entire evaluation code so that the public version of our results can be reproduced. Additionally, we commit to open sourcing the entire benchmark when either 1) the top open source models score over 95% on GSM1k or 2) June 2025, whichever comes earlier. See Section 3 for precise release criteria.

2 Related Work

A major inspiration of this work was the celebrated study on overfitting done on ImageNet classifiers in 2019 (Recht et al. [2019]). This work measured overfitting in ImageNet by creating new versions of CIFAR10 and ImageNet and measuring the performance gap between the public test set and the newly created sets they constructed. In this work, we do a similar analysis on GSM8k, one of the leading benchmarks for mathematical reasoning. GSM1k is modelled after the GSM8k dataset (Cobbe et al. [2021]), released by OpenAI in 2021, which consists of 8.5k grade school math problems. Each problem is designed to be solvable using only basic arithmetic operations ($+$, $-$, $\times$, $\div$) with a difficulty level appropriate for grade school students. As of June 2024, top models report benchmark accuracies of over 95% (Team et al. [2024]). Other popular benchmarks for reasoning include MATH (Hendrycks et al. [2021b]) , MMLU (Hendrycks et al. [2021a]), GPQA (Rein et al. [2023]).

2.1 Data Contamination

Because data contamination is a well known issue in the field (Balloccu et al. [2024], Magar and Schwartz [2022], Sainz et al. [2023], Jacovi et al. [2023], Xu et al. [2024]), model builders will frequently take great pains to minimize the likelihood of data contamination. For example, it is common to remove all data with too high of an n-gram overlap with the benchmark data (Brown et al. [2020]). Additionally, methods such as using embedding similarity attempt to remove all contaminated data that is too similar in embedding space to the dataset (Shi et al. [2024]).

Xu et al. [2024] propose using similar variants of a benchmark questions to detect if models favor the original wording as a proxy for data contamination. Srivastava et al. [2024] propose functional evaluations, where benchmarks are written in the form of functions that can generate an infinite number of specific evaluation datapoints, each with slightly different numbers. In this setup, whenever a language model is evaluated, functional evaluations generate a specific problem instance to evaluate the model on, which is then never used again. This reduces the worry of data contamination by ensuring that no datapoint is ever used twice. Like ours, their results indicate the LLMs may be severely overfit on benchmark data. The main advantage of our approach over a purely function based evaluation is that functional evaluations can only generate a tiny portion of the full problem space by producing variations of the same problem with slightly different numerical values. Their results also suggest substantial amounts of data contamination, including for frontier models, in the MATH dataset.

https://doi.org/10.52202/079017-1485

3 GSM1k

GSM1k consists of 1205 problems requiring only elementary mathematical reasoning to solve. We created GSM1k using human annotators. Annotators were prompted with 3 example GSM8k problems and asked to produce novel problems of a similar difficulty level. The precise instructions and UI given to the annotators is available in Appendix A. All problem annotators were instructed to create problems solvable with only basic arithmetic (addition, subtraction, multiplication, and division) and which did not require any advanced math concepts. As is the case with GSM8k, all problem solutions are positive integers[2]. No language models were used to construct this dataset.

To prevent data contamination concerns with GSM1k, we do not intend to release the dataset publicly at this time. However, we commit to releasing the full GSM1k dataset when at least one of the two following conditions have passed, whichever comes earlier. 1) Three open-source models with different pre-trained foundational model lineages reach 95% accuracy on GSM1k. 2) June 2025. At such a point, we believe that grade school mathematics will likely no longer be difficult enough to materially benchmark model releases and commit to releasing all data into the public domain under the MIT license. Additionally, to evaluate proprietary models, we were required to send over the dataset via API. Our belief is that model providers typically do not use such datapoints for model training. Nevertheless, in case GSM1k data is leaked through such means, we also hold out a small number of data points that have passed all quality checks but do not appear in the final GSM1k dataset. This data will also be released alongside GSM1k upon final release. We encourage future benchmarks to follow a similar pattern, where they are not released publicly lest they be gamed, but are precommitted to be released at a future date or upon a future condition. As part of this release, we will also open source our evaluation framework, which is based off of a fork of the LM Evaluation Harness by EleutherAI (Gao et al. [2023a]).

Finally, while we undertook extensive efforts to ensure maximum similarity between GSM8k and GSM1k, these results are only an approximation of an ideal world in which the test set of GSM8k was instead not publicly released and used for evaluations. We would recommend reading all results with the understanding that GSM8k and GSM1k are only highly similar, but not identically distributed despite all our efforts below.

3.1 Quality Checks

All questions passed through a total of 3 review layers. After initial creation, each task was manually reviewed by a subset of trusted annotators selected for strong past performance. These reviewers checked both for correctness as well as ensuring problems contained only grade school level math and proper formatting. To ensure that questions were answered correctly, we also do a second review layer by having an independent set of data annotators solve each question *without seeing the intended solution*. If this second solve produced a different answer to that of the initial solve, we discarded the problem. Finally, all problems were reviewed by a special team within Scale responsible for conducting general quality audits for data production. Out of a total of 2108 initial problems, 1419 passed the second solve stage and 1375 passed the general quality audit.

3.2 Matching the Difficulty Distribution of GSM8k

One important axis of recreating a benchmark is ensuring that new problems have a comparable difficulty to the original benchmark. To construct problems of difficulty N, we requested annotators to construct problems with N required resolution steps and prompted them with 3 examples from GSM8k with estimated difficulty N. The distribution of problems requested from annotators matched the estimated distribution in GSM8k. Difficulty is tricky to measure precisely, so we used an estimate based on the number of operations needed to solve the problem. This was extracted programmatically by counting the number of "calculator" tags in the problem solution. However, as not all problem solutions were formatted consistently, this estimate is only a rough estimate of actual difficulty.

[2]GSM8k has a few problems, likely errors, for which this is not the case.

Additionally, the number of resolution steps in a problem does not necessarily directly correlate with the true level of problem difficulty.

Past work has also found that LLMs struggle with problems with larger numbers (Gao et al. [2023b]) even if they can solve otherwise identical problems with smaller numbers. To remove this as a potential confounding variable, our final processing step is to discard candidate problems from GSM1k so that the answer magnitude distributions of GSM8k and GSM1k are as similar as possible. This selection process is described in Figure 3. GSM1k consists of the 1205 problems that survive this final winnowing. Additionally, we run several checks to ensure that our efforts to match benchmark

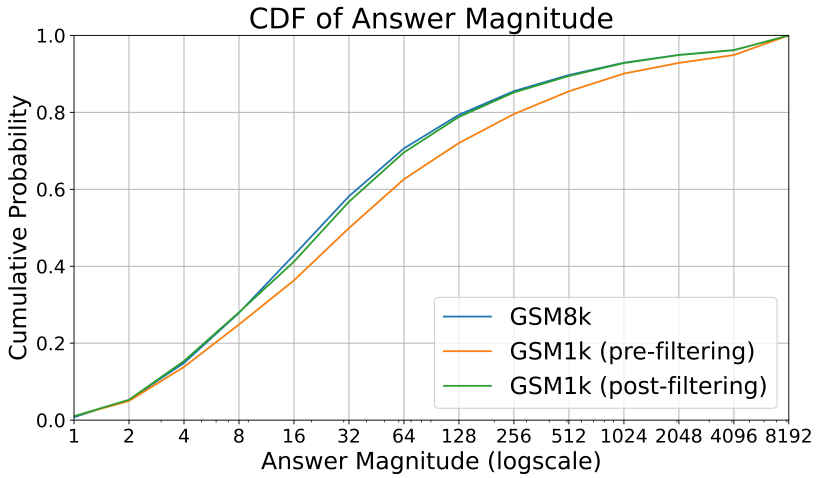

Figure 3: As the final step, we select 1205 problems to match the answer magnitude distribution of GSM8k as much as possible. The remaining problems are discarded and not included in the final dataset. Before discarding, we find that our generated problems tend to have slightly larger answers.

difficulty were successful.

3.2.1 Human Differentiation Rates

The first test we run is human distinguishability. We present human annotators with a set of five questions, four of which were randomly selected from the original GSM8k dataset and one of which was selected from the newly created GSM1k dataset, and rewarded annotators for finding the odd one out. In an audit conducted using 19 annotators who were not involved in the problem creation process, we found that annotators were able to correctly identify the lone GSM1k example 21.83% of the time out of 1205 attempts (20% is pure chance). Separately, we also tested several paper authors who had not yet seen the data and they were also unable to perform much better than random. This suggests minimal differences between GSM8k and GSM1k, at least as measured by the human eye.

3.2.2 Human Solve Rates

To ensure similar solve rates, we also asked annotators to solve questions under time pressure. 14 annotators who had not participated in the problem creation process attempted to solve as many GSM8k problems as they could in 15 minutes and were rewarded based on the number of problems they solved. We repeated this exact setup for GSM1k. Annotators were able to solve an average of 4.07 ± 0.93 problems on the GSM8k dataset. They were able to solve 4.36 ± 1.11 problems on the GSM1k dataset, where the error rates are the standard deviations of the evaluation. This suggests that GSM1k is comparable in difficulty (and perhaps even slightly easier) than GSM8k. As such, substantial decreases in model accuracy on GSM1k compared to GSM8k are likely not explainable due to differences in dataset difficulty.

https://doi.org/10.52202/079017-1485

3.2.3 LLM Solve Rates

Finally, we sanity check our results by measuring solve rates of several models that are known to not be contaminated by GSM8k due to being released before the publication of the GSM8k dataset. Due to the relative scarcity of LLMs trained only on pre-2021 data, we evaluate only GPT-NeoX-20B (Black et al. [2022]) and GPT-2 (Radford et al. [2019]). For these two language models, we find minimal difference between their solve rates of GSM8k and GSM1k (Figure 12).

4 Results

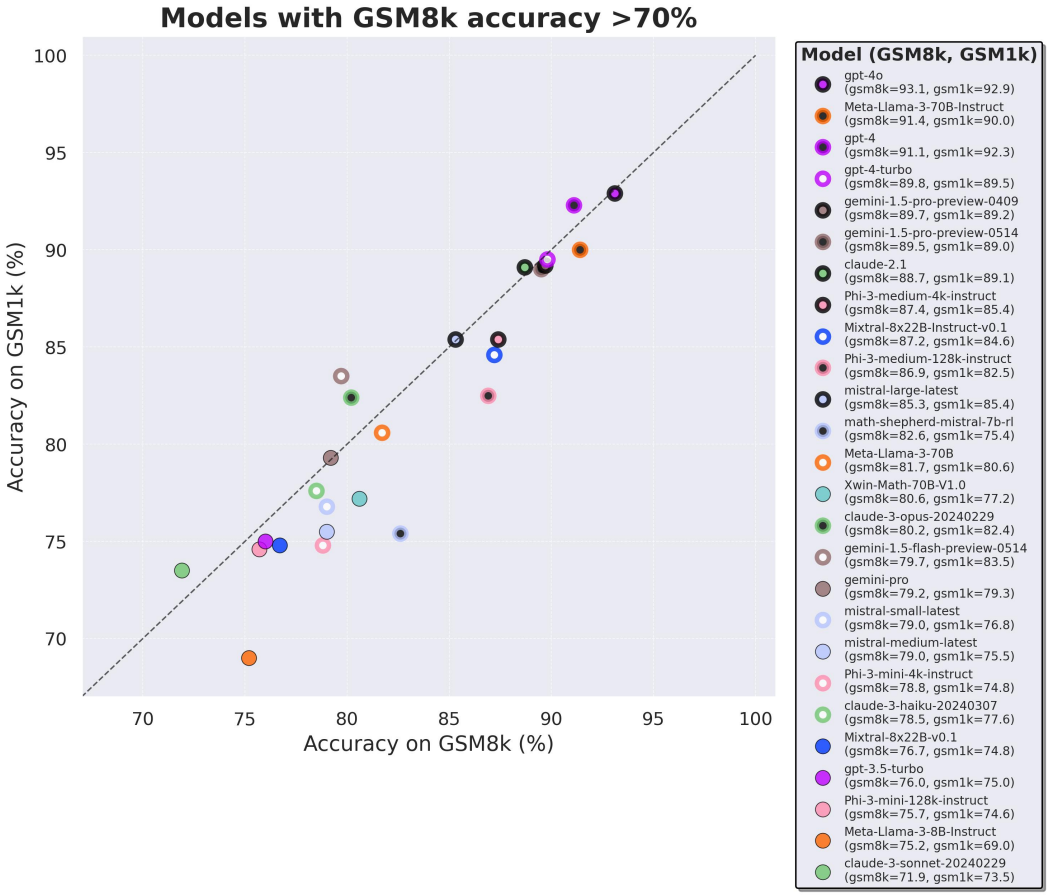

Figure 4: Models with over 70% accuracy on GSM8k compared to the line of no overfit. This plot is zoomed into the relevant sections (70-100% accuracy). Note that some models, especially the Claude family, perform above the 45 degree line, which is consistent with our findings in Section 3 that GSM1k is slightly easier than GSM8k. In contrast, many other models lie well below this line.

To evaluate models, we use a fork of EleutherAI's LM Evaluation Harness with minor modifications. We use the default settings for evaluation, except for increasing the maximum number of allowed generated tokens from 256 to 1000, as we notice that the default setting did not allow some models to complete their full chain-of-thought reasoning before being truncated. Both GSM8k and GSM1k questions are run with the same prompt of using 5 randomly drawn examples from the GSM8k train set, as is standard in the field. An example prompt is provided in Appendix B. All open-source models are evaluated at temperature 0 for reproducibility. For open source models, we use vLLM to speed up model inference if a model is compatible with the library. Otherwise, we default to inference using standard HuggingFace libraries. Closed-source models were queried through the LiteLLM library

which unifies the API call format for all proprietary models evaluated. All API model results were from queries between April 16 - June 5, 2024 and use the default settings.

LM Evaluation Harness uses an automatic evaluation method which extracts the last numeric answer in the response and compares this to the correct answer. However, in some cases, models will produce "correct" answers in a format that do not match the given examples, resulting in their answers being marked as incorrect. To explore the effect of this on the results, we run an ablation where we select a subset of models and use human annotation to manually extract answers that are not correctly formatted (Appendix H). We do not find major changes in our findings for the models examined.

As model benchmark performance is highly dependent on choice of prompt and evaluation setting, our reported GSM8k numbers may occasionally be below the reported model benchmark numbers, as we use a standardized setting for all models instead of the prompt that maximizes each individual model's performance. Additionally, we explore the effect of different prompt formulations with several ablations. In Appendix C, we report results with an alternative prompting format that uses non-GSM8k examples as n-shot examples and a slightly different answer phrasing. Additionally, we explore the effect of varying the number and source of the n-shot examples used in Appendix I and J. While the precise benchmark accuracies vary depending on the setup, we find that the general trends of overfitting hold consistently across our ablations. We will release the full evaluation code for transparency.

In addition to evaluating widely known models, we additionally evaluate several lesser known models that sit near the top of the OpenLLMLeaderboard and discover evidence of Goodhart's law: many of these models perform substantially worse on GSM1k, suggesting that they are primarily gaming the GSM8k benchmark rather than improving model reasoning capabilities. The full set of results, including the performance table for all models, can be found in Appendix D. For fair comparison, we partition the models by performance on GSM8k and compare them to other models which perform similarly (Figures 4, 11, 12).

5 Analysis

The interpretation of evaluation results, like the interpretations of dreams, is often a very subjective endeavor. While we report our objective results in Section 4 and Appendix D, here we describe four major takeaways from interpreting the results in a more subjective manner.

5.1 Lesson 1: Some Model Families are Systematically Overfit

While it is difficult to draw conclusions from singular data points or model releases, examining a family of models and observing a pattern of overfitting enables us to make more definitive statements. Several families of models, including the Phi and Mistral families of models, show systematic tendencies to perform stronger on GSM8k compared to GSM1k for almost every release and scale of models. Other model families, such as Yi, Xwin, Gemma and CodeLlama also show this pattern to a lesser extent.

5.2 Lesson 2: Other Models, Especially Frontier Models, Show No Signs of Overfitting

Nevertheless, we find that many models, through all regions of performance, show minimal signs of being overfit. In particular, we find that all frontier or close-to-frontier models (including the proprietary Mistral Large) appear to perform similarly on both GSM8k and GSM1k. We posit two potential hypotheses for this: 1) frontier models have sufficiently advanced reasoning capability so that they can generalize to new problems even if they have already seen GSM8k problems in their training set, 2) frontier model builders may be more careful about data contamination.

While it is impossible to know for certain without looking at the training set for each model, one piece of evidence in favor of the former is that Mistral Large is the *only* model in the Mistral family to show no signs of overfitting. Since the hypothesis that Mistral took unique care in ensuring that

https://doi.org/10.52202/079017-1485

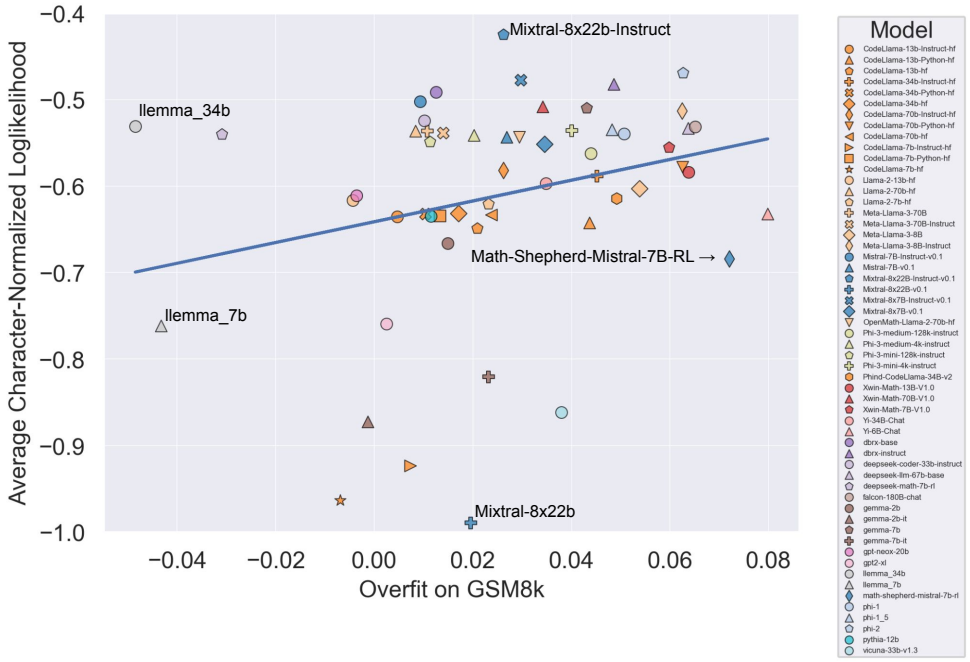

Figure 5: Comparison between overfit on GSM8k (x-axis) and average sequence-level log-likelihood on the GSM8k test set (y-axis). We find that there is a correlation between overfit on GSM8k and sequence-level log-likelihood, suggesting that, in general, models that have a high overfit generally have a higher probability of generating the test set. This suggests that some of the GSM8k test set may have leaked into the model training data. The line of best fit is in blue. Additionally, we highlight 5 "outlier" models which we discuss further with Lesson 4.

only their largest model was free from data contamination seems unlikely, we lean instead towards the hypothesis that sufficiently strong LLMs also learn elementary reasoning ability during training. If a model learns strong enough reasoning capabilities to solve problems of a given difficulty, it will be able to generalize to new problems even if GSM8k has appeared in its training set.

5.3 Lesson 3: Overfit Models Are Still Capable of Reasoning

One worry about model overfitting is that models are incapable of reasoning and are only memorizing answers seen in the training data. Our results do not support this conjecture. The fact that a model is overfit does not mean that it is poor at reasoning, merely that it is not as good as the benchmarks might indicate it to be. In fact, we find that many of the most overfit models are still capable of reasoning and solving novel problems. For example, while Phi-2 has a 6% drop in accuracy between GSM8k and GSM1k, we find that it is still able to correctly solve over half of GSM1k problems – which are certain to not have appeared in its training distribution. This performance is similar to that of much larger models such as Llama2-70B, which contains over 25x as many parameters. Similarly, Mistral models remain some of the strongest open source models, even accounting for their overfitting. This provides additional evidence for our lesson that sufficiently strong models learn elementary reasoning, even if benchmark data accidentally leaked into the training distribution, as is likely to be the case for the most overfit models.

5.4 Lesson 4: Data Contamination Is Likely Not The Full Explanation for Overfitting

A priori, a natural hypothesis is that the primary cause for overfitting is data contamination, e.g. that the test set was leaked in the pre-training or instruction fine-tuning part of the model creation. Previous work has suggested that models put higher log-likelihoods on data that they have seen

during training (Carlini et al. [2023]). We test the hypothesis that data contamination is the cause of overfitting by measuring a model's probability of generating an example from the GSM8k test set and comparing it to how overfit it is on GSM8k vs GSM1k, using the assumption that a model's probability of generating the GSM8k test set is a proxy for whether the sequence is likely to have appeared in the training set. We normalize by c, the number of characters in the sequence, to make the log-likelihood calculations comparable between sequences and models with different tokenizers. Formally, we have:

$$\frac{1}{c}\sum_{i}\log p(x_i|x_{<i}) \tag{1}$$

with c being the number of characters in the sequence. Figure 5 shows the result of this plot against the gap between GSM8k and GSM1k performance. We indeed find a positive relationship between the two values. We observe a Spearman's rank correlation of 0.36 between the per-character log-likelihood of generating GSM8k and the performance gap between GSM8k and GSM1k ($p = 0.03$), and the relationship suggests that every percentage point difference in GSM8k and GSM1k performance is associated with an increase of 1.2×10^{-2} in the per-character log-likelihood. This result suggests that some of the reason for overfitting is due to partial memorization of the test set. For completeness, we also report the standard Pearson $r^2 = 0.26$ and the Kendall's τ of 0.29, but note that Pearson r^2 is not the ideal metric due to the curve-of-best-fit not appearing linear.

Nevertheless, data contamination is likely not the full story. We observe this via the presence of several outliers, which cause the $r^2 = 0.36$ value to be relatively low. Examining these outliers carefully reveals that the model with the lowest per-character log-likelihood (Mixtral-8x22b) and the model with the highest per-character log-likelihood (Mixtral-8x22b-Instruct) are not only variations of the same model, but also have similar levels of overfit (Jiang et al. [2024]). Perhaps more intriguingly, one the most overfit models we discovered (Math-Shepherd-Mistral-7B-RL (Yu et al. [2023])) had a relatively low per-character log-likelihood. Math Shepherd trains a reward model on process level data using synthetic data. As such, we hypothesize that the reward modelling process may have leaked information about the correct reasoning chains for GSM8k even if the problems themselves did not ever appear in the dataset. Finally, we observe that the Llema models (Azerbayev et al. [2024]) have both high log-likelihoods and minimal overfit. These models are open-sourced alongside their training data, and the authors report finding a very small number of GSM8k examples in the training corpus. Nevertheless, they also find (and our study supports) that these few instances do not lead to overfitting. The existence of these outliers suggests that overfitting on GSM8k is not purely due to data contamination, but rather may be through other indirect means, such as model builders collecting data similar in nature to benchmarks as training data or selecting final model checkpoints based on performance on benchmarks, even if the model itself may have not seen the GSM8k dataset at any point via training. Conversely, the reverse is also true: small amounts of data contamination do not necessarily lead to overfitting.

6 Discussion

We create GSM1k, a novel dataset designed to measure LLM overfitting on GSM8k. When benchmarking leading open- and closed-source models, we find substantial evidence that many models have been contaminated by benchmark data, with models showing performance drops of up to 8% accuracy. Additionally, we find that several model families show consistent overfitting across almost all model sizes and versions. An extended analysis reveals a positive relationship between a model's likelihood of generating data points in GSM8k and its performance difference between GSM8k and GSM1k, suggesting evidence of data contamination as one of the underlying causes. Nevertheless, we find that frontier models exhibit little to no evidence of overfitting and that many models, even the most heavily overfit families, show strong signs of generalizable mathematical reasoning.

https://doi.org/10.52202/079017-1485

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, April 2024. URL http://arxiv.org/abs/2404.14219. arXiv:2404.14219 [cs].

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program Synthesis with Large Language Models, August 2021. URL http://arxiv.org/abs/2108.07732. arXiv:2108.07732 [cs].

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An Open Language Model For Mathematics, March 2024. URL http://arxiv.org/abs/2310.10631. arXiv:2310.10631 [cs].

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs, February 2024. URL http://arxiv.org/abs/2402.03927. arXiv:2402.03927 [cs].

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An Open-Source Autoregressive Language Model, April 2022. URL http://arxiv.org/abs/2204.06745. arXiv:2204.06745 [cs].

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying Memorization Across Neural Language Models, March 2023. URL http://arxiv.org/abs/2202.07646. arXiv:2202.07646 [cs].

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian,

Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code, July 2021. URL http://arxiv.org/abs/2107.03374. arXiv:2107.03374 [cs].

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, November 2021. URL http://arxiv.org/abs/2110.14168. arXiv:2110.14168 [cs].

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, December 2023a. URL https://zenodo.org/records/10256836. tex.version: v0.4.0.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided Language Models, January 2023b. URL http://arxiv.org/abs/2211.10435. arXiv:2211.10435 [cs].

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks Are All You Need, October 2023. URL http://arxiv.org/abs/2306.11644. arXiv:2306.11644 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021a. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

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*, 2021b.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.308. URL https://aclanthology.org/2023.emnlp-main.308.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao,

 https://doi.org/10.52202/079017-1485

Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, January 2024. URL http://arxiv.org/abs/2401.04088. arXiv:2401.04088 [cs].

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues?, April 2024. URL http://arxiv.org/abs/2310.06770. arXiv:2310.06770 [cs].

Inbal Magar and Roy Schwartz. Data Contamination: From Memorization to Exploitation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.18. URL https://aclanthology.org/2022.acl-short.18.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben

Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24, 2019.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet?, June 2019. URL http://arxiv.org/abs/1902.10811. arXiv:1902.10811 [cs, stat].

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark, November 2023. URL https://arxiv.org/abs/2311.12022v1.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting Pretraining Data from Large Language Models, March 2024. URL http://arxiv.org/abs/2310.16789. arXiv:2310.16789 [cs].

Saurabh Srivastava, Annarose M. B, Anto P V, Shashank Menon, Ajay Sukumar, Adwaith Samod T, Alan Philipose, Stevin Prince, and Sooraj Thomas. Functional Benchmarks for Robust Evaluation of Reasoning Performance, and the Reasoning Gap, February 2024. URL http://arxiv.org/abs/2402.19450. arXiv:2402.19450 [cs].

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

 https://doi.org/10.52202/079017-1485

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 Kępa, 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, Z. J. 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

 https://doi.org/10.52202/079017-1485

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, T. J. 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, M. K. Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, 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, 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, April 2024. URL http://arxiv.org/abs/2312.11805. arXiv:2312.11805 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023a. URL http://arxiv.org/abs/2302.13971. arXiv:2302.13971 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023b. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

 https://doi.org/10.52202/079017-1485

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking Benchmark Leakage in Large Language Models, April 2024. URL http://arxiv.org/abs/2404.18824. arXiv:2404.18824 [cs].

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models, October 2023. URL http://arxiv.org/abs/2309.12284. arXiv:2309.12284 [cs].