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 †

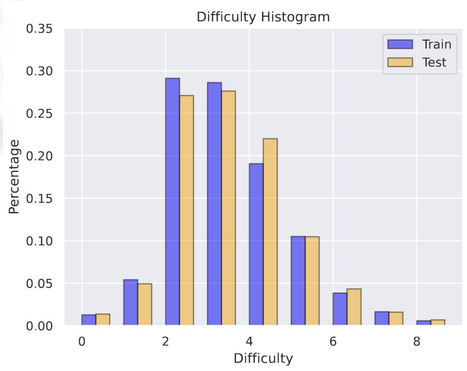
Motivation

- Proper benchmarking of current LLM abilities is crucial for ensuring progress continues in the right direction.
- Since LLMs are trained on large corpora of data scraped from the internet, there are concerns that current benchmarks may include examples resembling questions in these benchmarks.
- This contamination may lead to models appearing to have stronger capabilities as they may simply repeat correct answers encountered during pre- or post-training.

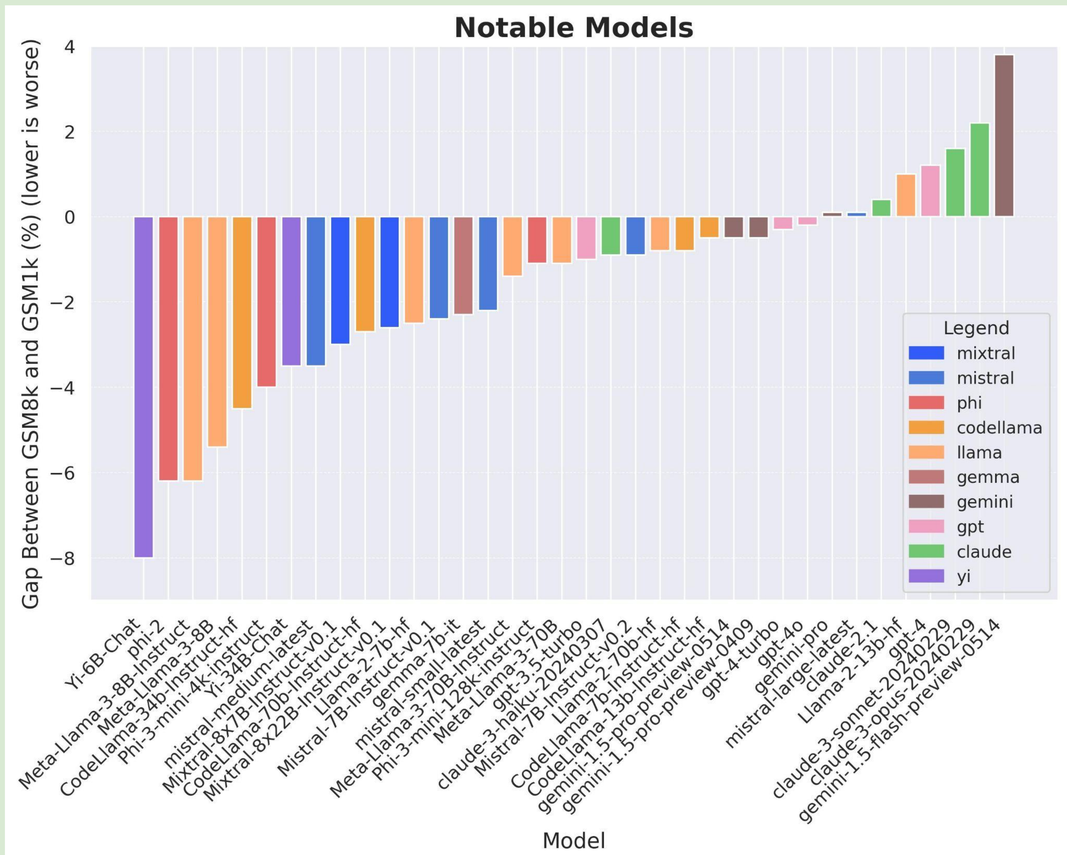
Creating GSM1k

GSM1k is a dataset of 1,205 problems requiring only elementary mathematical reasoning. Human annotators created the problems after being shown example GSM8K problems and instructed to produce novel problems of similar difficulty.

- Human distinguishability test: We presented human annotators with a set of five questions, four of which were randomly selected from GSM8k dataset and one from GSK1k.
- We found that annotators were able to identify the GSM1k example 21.83% of the time out of 1205 attempts (20% is pure chance).
- This suggests minimal differences between GSM8k and GSM1k, at least as measured by the human eye.

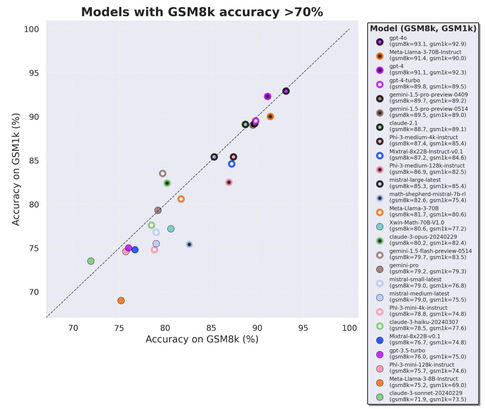


GSM1k is an uncontaminated replication of the GSM8k dataset – used to investigate model benchmark performance.



View the code and paper!

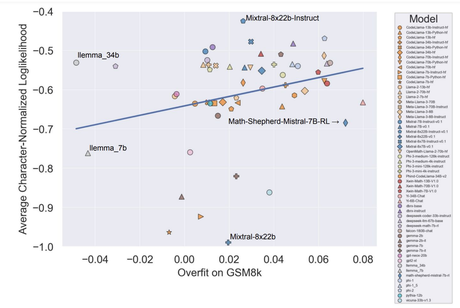
Results



Some Model Families are Systematically Overfit. Several families of models show systematic tendencies to perform stronger on GSM8k compared to GSM1k for almost model type

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 appear to perform similarly on both GSM8k and GSM1k.

Overfit Models Are Still Capable of Reasoning. 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.



Data Contamination Is Likely Not The Full Explanation for Overfitting. We test the hypothesis that data contamination is the cause of overfitting.

- We measure a model's probability of generating an example from the GSM8k test set. We find a Pearson r^2=0.26 and the Kendall's τ of 0.29, suggesting moderate correlation.
- Overfitting on GSM8k is not purely due to data contamination, but rather may be through other indirect means.