A   Checklist

1. For all authors...

    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

    (b) Did you describe the limitations of your work? [Yes]

    (c) Did you discuss any potential negative societal impacts of your work? [N/A]

    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

    (a) Did you state the full set of assumptions of all theoretical results? [N/A]

    (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...

    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] included in supplement, will open source shortly

    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Appendix F includes standard errors and statistical significance.

    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We run all models on a cluster with 8 x A100 nodes. Most models complete evaluation within a few minutes.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

    (a) If your work uses existing assets, did you cite the creators? [N/A]

    (b) Did you mention the license of the assets? [N/A]

    (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] GSM1k is yet unreleased, with a future release date. We provide the full dataset here. We ask that reviewers refrain from sharing this dataset publicly.

    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] All problems created by annotators hired by Scale AI.

    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix C

    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] Annotators were paid 20-25 / hour, depending on performance, experience, and bonus incentives. In total, Scale paid out around 180K to human annotators to create this benchmark.

19

B Dataset Documentation

1. **Construction.** GSM1k is a dataset of 1205 questions requiring elementary mathematical reasoning to solve. All problems are intended to be solvable using only the four basic arithmetic operators.

2. **Creation.** GSM1k was created using human annotation from scratch without any usage of LLMs. Human annotators were hired by Scale AI and paid between 20 and 25 dollars per hour. All annotators were based in the United States. In total, this dataset paid out around $180,000 dollars to human annotators, including costs resulting from problem creation and solving, quality assurance checks, as well as experiments done to compare the difficulty distribution with GSM8k.

3. **Intent.** This dataset is intended to be used as a held-out version of GSM8k to measure data contamination. As such, it largely mimics the format and style of GSM8k. All answers are a non-negative integer.

4. **Release.** Our dataset is not published at this time, to prevent risk of data contamination in future models. We will release Croissant metadata when the dataset is public, with the conditions described in the main paper.

5. **Liability.** The authors bear all responsibility in case of violation of rights. Due to Scale AI commissioning the construction of this dataset from scratch primarily for this purpose of this paper, we do not anticipate any copyright or other issues. The dataset (yet unreleased) will be released with the MIT license.

6. **Preservation.** We plan to release the full dataset on Github as well as HuggingFace so it remains publicly accessible to anyone who wishes to use it. The formatting will be 1205 rows with a question and answer column.

20

C   Annotator Instructions

We provide the annotator instructions given below.

```
Welcome to the Grade School Math Question Development project. The goal
    of this project is to create questions and answers similar to what is
     found in an 8th—grade math quiz. Our goal is to develop high—quality
     questions that are almost the same as what is found in the dataset
    but are entirely unique. You will see three example questions and
    their corresponding answers in each task. These examples will guide
    you to create completely new questions and answers. It's important to
     note that you cannot use chatbots or language models to help you
    develop these Q&A pairs. You may be removed from the project if we
    detect any use of chatbots. Crucially, your Q&A pairs must be
    original creations and cannot be paraphrased versions of the examples
     .

Your workflow for this project will be as follows:

Review the examples: In each task you will be shown examples from an 8th—
    grade question—and—answer dataset. Review the examples to inform how
    you can create your question and answer pair.

Problem Creation: Problems should follow step guidance in the task. Don't
     reuse a problem setting. If you wrote a problem about Rogers trip to
     the grocery store, don't write another problem using the same
    premise. All questions should have a resolution of 1 or higher. We do
     not want any questions with a negative integer or zero as the answer
     .

Craft the resolution steps: Calculations should be simple enough an 8th
    grader can complete with a pen and paper. Only use elementary
    arithmetic operations (addition, subtraction, multiplication,
    division)

Provide the final Answer: Answers should be a single integer value. Any
    units should be specified as part of the question (e.g. "How much
    money, in dollars, does Robert have?"). Simple decimal numbers (e.g.
    3.25) can be part of the intermediate steps in the problem, but final
     answers should always be integers.

Check your work: We will utilize quality control process to ensure
    accuracy but it is crucial to check your work!
```

21

Review the provided examples
Read the instructions carefully before continuing.

Please review the instructions carefully before continuing.

Write your math problem here:
..

- Problems should follow step guidance in the task
- Calculations should be simple enough to do in your head
- Answers should be a single integer value. Any units should be specified as part of the question (e.g. "How much money, in dollars, does Robert have?"). Simple decimal numbers (e.g. 3.25) can be part of the intermediate steps in the problem, but final answers should always be integers.
- Only use elementary arithmetic operations
- Don't reuse a problem setting. If you wrote a problem about Rogers trip to the grocery store, don't write another problem using the same premise.

Start typing here...

0 words

Press `Shift` + `Enter` to submit your message.

Submit Message

Figure 6: What annotators saw before seeing three example prompts drawn from GSM8k.

22

D   N-shot Prompt (examples selected randomly from GSM8k train)

Below is an example prompt. For each question, we select five random examples from GSM8k to use as n-shot examples, which vary for each new question from the GSM1k/GSM8k test set. While evaluation methods vary between models, this is the most common approach to evaluating GSM8k.

```
Question: Jen and Tyler are gymnasts practicing flips. Jen is practicing the triple-flip
while Tyler is practicing the double-flip. Jen did sixteen triple-flips during practice.
Tyler flipped in the air half the number of times Jen did. How many double-flips did Tyler do?
Answer: Jen did 16 triple-flips, so she did 16 * 3 = <<16*3=48>>48 flips.
Tyler did half the number of flips, so he did 48 / 2 = <<48/2=24>>24 flips.
A double flip has two flips, so Tyler did 24 / 2 = <<24/2=12>>12 double-flips.
#### 12
Question: Four people in a law firm are planning a party. Mary will buy a platter of pasta
for $20 and a loaf of bread for $2. Elle and Andrea will split the cost for buying 4 cans
of soda which cost $1.50 each, and chicken wings for $10. Joe will buy a cake that costs
$5. How much more will Mary spend than the rest of the firm put together?
Answer: Mary will spend $20 + $2 = $<<20+2=22>>22.
Elle and Andrea will spend $1.5 x 4 = $<<1.5*4=6>>6 for the soda.
Elle and Andrea will spend $6 + $10 = $<<6+10=16>>16 for the soda and chicken wings.
Elle, Andrea, and Joe together will spend $16 + $5 = $<<16+5=21>>21.
So, Mary will spend $22 - $21 = $<<22-21=1>>1 more than all of them combined.
#### 1
Question: A charcoal grill burns fifteen coals to ash every twenty minutes of grilling.
The grill ran for long enough to burn three bags of coals. Each bag of coal contains 60
coals. How long did the grill run?
Answer: The grill burned 3 * 60 = <<3*60=180>>180 coals.
It takes 20 minutes to burn 15 coals, so the grill ran for 180 / 15 * 20 =
<<180/15*20=240>>240 minutes.
#### 240
Question: A bear is preparing to hibernate for the winter and needs to gain 1000 pounds.
At the end of summer, the bear feasts on berries and small woodland animals. During autumn,
it devours acorns and salmon. It gained a fifth of the weight it needed from berries during
summer, and during autumn, it gained twice that amount from acorns. Salmon made up half of
the remaining weight it had needed to gain. How many pounds did it gain eating small animals?
Answer: The bear gained 1 / 5 * 1000 = <<1/5*1000=200>>200 pounds from berries.
It gained 2 * 200 = <<2*200=400>>400 pounds from acorns.
It still needed 1000 - 200 - 400 = <<1000-200-400=400>>400 pounds.
Thus, it gained 400 / 2 = <<400/2=200>>200 pounds from salmon.
Therefore, the bear gained 400 - 200 = <<400-200=200>>200 pounds from small animals.
#### 200
Question: Brendan can cut 8 yards of grass per day, he bought a lawnmower and it helped
him to cut more yards by Fifty percent per day. How many yards will Brendan be able to cut
after a week?
Answer: The additional yard Brendan can cut after buying the lawnmower is 8 x 0.50 =
<<8*0.50=4>>4 yards.
So, the total yards he can cut with the lawnmower is 8 + 4 = <<8+4=12>>12.
Therefore, the total number of yards he can cut in a week is 12 x 7 = <<12*7=84>>84 yards.
#### 84
```

23

E Results with an Alternative Prompt

As an ablation, we evaluate all models with an alternative prompt scheme and compare results with our primary findings. This prompt is available under the LM Evaluation Harness as a "chain-of-thought" prompt. However, manually examining the prompt (provided in full below) reveals that the primary difference with the standard n-shot prompt lies not in chain-of-thought reasoning but rather using a set of non-GSM8k problems as guiding examples as well as providing an alternative answer format. We choose to use the standard prompt to match typical evaluation methods widespread in the field but also report these results for completeness.

```
Q: There are 15 trees in the grove. Grove workers will plant trees in the
    grove today. After they are done, there will be 21 trees. How many
    trees did the grove workers plant today?
A: There are 15 trees originally. Then there were 21 trees after some
    more were planted. So there must have been 21 − 15 = 6. The answer is
    6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how
    many cars are in the parking lot?
A: There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. The answer
    is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many
    pieces do they have left in total?
A: Originally, Leah had 32 chocolates. Her sister had 42. So in total
    they had 32 + 42 = 74. After eating 35, they had 74 − 35 = 39. The
    answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12
    lollipops. How many lollipops did Jason give to Denny?
A: Jason started with 20 lollipops. Then he had 12 after giving some to
    Denny. So he gave Denny 20 − 12 = 8. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom
    and dad. How many toys does he have now?
A: Shawn started with 5 toys. If he got 2 toys each from his mom and dad,
    then that is 4 more toys. 5 + 4 = 9. The answer is 9.

Q: There were nine computers in the server room. Five more computers were
    installed each day, from monday to thursday. How many computers are
    now in the server room?
A: There were originally 9 computers. For each of 4 days, 5 more
    computers were added. So 5 ∗ 4 = 20 computers were added. 9 + 20 is
    29. The answer is 29.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On
    wednesday, he lost 2 more. How many golf balls did he have at the end
    of wednesday?
A: Michael started with 58 golf balls. After losing 23 on tuesday, he had
    58 − 23 = 35. After losing 2 more, he had 35 − 2 = 33 golf balls.
    The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money
    does she have left?
A: Olivia had 23 dollars. 5 bagels for 3 dollars each will be 5 x 3 = 15
    dollars. So she has 23 − 15 dollars left. 23 − 15 is 8. The answer is
    8.
```

We report our results in Table F. While there is significant variance based on prompt, the general trend of which model families are overfit is similar.

24
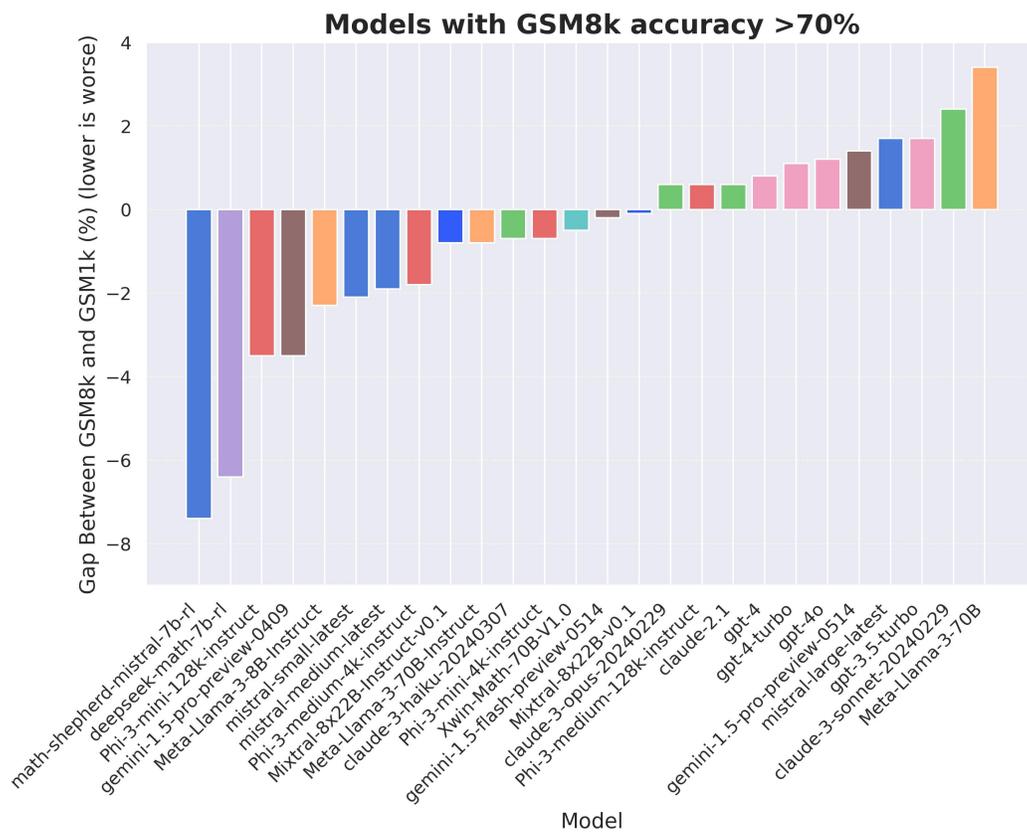
Figure 7: Gap in accuracy between GSM8k and GSM1k for models that score above 70% on GSM8k.

25

F   Results Table

We report our full results in Table F. Models are sorted by the difference in performance between GSM8k and GSM1k. Because all models are evaluated using the standard LM Evaluation Harness prompt and evaluation format, model performance on GSM8k may not match reported benchmark numbers. In particular, answers that do not match the 5-shot example format are marked incorrect even if they are otherwise "correct." Our focus is primarily on the difference between GSM8k and GSM1k performance, holding evaluation setting constant. The Z-score and p-value are calculated for a two-tailed two proportion Z-test. Alternative prompt results are also included. For details, see Appendix E.

Standard Prompt

Model	Diff	GSM8k	GSM1k	Z-score	p-value
Yi-6B-Chat	0.080	0.437	0.357	4.135	0.000
math-shepherd-mistral-7b-rl	0.072	0.826	0.754	4.488	0.000
command	0.065	0.447	0.383	3.336	0.000
Xwin-Math-13B-V1.0	0.064	0.660	0.596	3.334	0.000
phi-2	0.063	0.566	0.504	3.167	0.001
Meta-Llama-3-8B-Instruct	0.062	0.752	0.690	3.532	0.000
Xwin-Math-7B-V1.0	0.060	0.552	0.492	3.040	0.001
Meta-Llama-3-8B	0.054	0.502	0.448	2.734	0.003
phi-1.5	0.051	0.324	0.274	2.814	0.002
Phind-CodeLlama-34B-v2	0.049	0.419	0.370	2.531	0.006
CodeLlama-34b-Instruct-hf	0.045	0.426	0.381	2.338	0.010
Phi-3-medium-128k-instruct	0.044	0.869	0.825	3.103	0.001
CodeLlama-13b-Python-hf	0.044	0.223	0.179	2.759	0.003
gemma-7b	0.043	0.519	0.476	2.198	0.014
Phi-3-mini-4k-instruct	0.040	0.788	0.748	2.385	0.009
Yi-34B-Chat	0.035	0.685	0.650	1.883	0.030
mistral-medium-latest	0.035	0.790	0.755	2.104	0.018
Mixtral-8x7B-v0.1	0.035	0.591	0.557	1.771	0.038
Xwin-Math-70B-V1.0	0.034	0.806	0.772	2.107	0.018
Mixtral-8x7B-Instruct-v0.1	0.030	0.660	0.630	1.588	0.056
Mistral-7B-v0.1	0.027	0.391	0.364	1.421	0.078
Mixtral-8x22B-Instruct-v0.1	0.026	0.872	0.846	1.913	0.028
CodeLlama-70b-Instruct-hf	0.026	0.513	0.486	1.323	0.093
Llama-2-7b-hf	0.025	0.141	0.116	1.892	0.029
Mistral-7B-Instruct-v0.1	0.025	0.353	0.329	1.309	0.095
CodeLlama-70b-hf	0.024	0.478	0.454	1.221	0.111
gemma-7b-it	0.023	0.325	0.302	1.247	0.106
mistral-small-latest	0.022	0.790	0.768	1.343	0.090
CodeLlama-13b-hf	0.021	0.236	0.215	1.247	0.106
Phi-3-medium-4k-instruct	0.020	0.874	0.854	1.519	0.064

26

Standard Prompt

Model	Diff	GSM8k	GSM1k	Z-score	p-value
Mixtral-8x22B-v0.1	0.020	0.767	0.748	1.138	0.127
CodeLlama-34b-hf	0.017	0.354	0.337	0.919	0.179
gemma-2b	0.015	0.185	0.170	0.966	0.167
Meta-Llama-3-70B-Instruct	0.014	0.914	0.900	1.251	0.105
CodeLlama-7b-Python-hf	0.013	0.131	0.118	1.040	0.149
pythia-12b	0.011	0.036	0.025	1.701	0.044
Phi-3-mini-128k-instruct	0.011	0.757	0.746	0.645	0.260
Meta-Llama-3-70B	0.011	0.817	0.806	0.707	0.240
CodeLlama-34b-Python-hf	0.010	0.312	0.301	0.549	0.291
gpt-3.5-turbo	0.009	0.760	0.750	0.546	0.293
Mistral-7B-Instruct-v0.2	0.009	0.428	0.419	0.469	0.319
claude-3-haiku-20240307	0.009	0.785	0.776	0.532	0.298
Llama-2-70b-hf	0.008	0.552	0.544	0.445	0.328
CodeLlama-7b-Instruct-hf	0.007	0.177	0.169	0.472	0.319
gemini-1.5-pro-preview-0514	0.006	0.895	0.890	0.472	0.318
gemini-1.5-pro-preview-0409	0.005	0.897	0.892	0.403	0.343
CodeLlama-13b-Instruct-hf	0.005	0.267	0.262	0.257	0.399
gpt-4-turbo	0.003	0.898	0.895	0.270	0.394
gpt2-xl	0.002	0.009	0.007	0.778	0.218
gpt-4o	0.002	0.931	0.929	0.219	0.413
gemini-pro	-0.001	0.792	0.793	-0.081	0.532
mistral-large-latest	-0.001	0.853	0.854	-0.049	0.519
gemma-2b-it	-0.001	0.111	0.112	-0.106	0.542
claude-2.1	-0.004	0.887	0.891	-0.336	0.632
CodeLlama-7b-hf	-0.007	0.126	0.133	-0.525	0.700
Llama-2-13b-hf	-0.011	0.236	0.246	-0.629	0.735
gpt-4	-0.012	0.911	0.923	-1.161	0.877
claude-3-sonnet-20240229	-0.016	0.719	0.735	-0.894	0.814
claude-3-opus-20240229	-0.022	0.802	0.824	-1.421	0.922
deepseek-math-7b-rl	-0.031	0.187	0.217	-1.963	0.975
gemini-1.5-flash-preview-0514	-0.038	0.797	0.835	-2.507	0.994

27

Alternative Prompt

Model	Diff	GSM8k	GSM1k	Z-score	p-value
math-shepherd-mistral-7b-rl	0.074	0.820	0.746	4.504	0.000
deepseek-math-7b-rl	0.064	0.760	0.696	3.672	0.000
Yi-6B-Chat	0.058	0.426	0.368	2.964	0.002
CodeLlama-34b-Python-hf	0.056	0.337	0.280	3.059	0.001
command	0.051	0.457	0.407	2.596	0.005
phi-1.5	0.050	0.321	0.271	2.786	0.003
Xwin-Math-13B-V1.0	0.042	0.662	0.620	2.212	0.013
CodeLlama-70b-Instruct-hf	0.040	0.529	0.489	2.047	0.020
CodeLlama-70b-hf	0.037	0.517	0.480	1.878	0.030
gemma-7b	0.036	0.568	0.532	1.826	0.034
gemini-1.5-pro-preview-0409	0.035	0.908	0.873	2.883	0.002
Phi-3-mini-128k-instruct	0.035	0.818	0.783	2.211	0.014
phi-2	0.034	0.552	0.518	1.744	0.041
Xwin-Math-7B-V1.0	0.033	0.530	0.497	1.680	0.046
CodeLlama-7b-hf	0.027	0.123	0.095	2.242	0.012
Mistral-7B-Instruct-v0.2	0.027	0.437	0.410	1.389	0.082
gemma-7b-it	0.023	0.254	0.231	1.394	0.082
Meta-Llama-3-8B-Instruct	0.023	0.774	0.751	1.362	0.087
CodeLlama-7b-Instruct-hf	0.022	0.187	0.165	1.493	0.068
Yi-34B-Chat	0.022	0.679	0.656	1.170	0.121
mistral-small-latest	0.021	0.782	0.761	1.305	0.096
CodeLlama-13b-Python-hf	0.021	0.218	0.197	1.299	0.097
CodeLlama-34b-hf	0.020	0.330	0.310	1.097	0.136
pythia-12b	0.019	0.049	0.030	2.553	0.005
mistral-medium-latest	0.019	0.789	0.770	1.152	0.125
Phi-3-medium-4k-instruct	0.018	0.901	0.883	1.428	0.077
Mixtral-8x7B-Instruct-v0.1	0.016	0.679	0.662	0.870	0.192
gemma-2b	0.016	0.194	0.178	1.020	0.154
Llama-2-7b-hf	0.014	0.142	0.128	1.021	0.154
Phind-CodeLlama-34B-v2	0.014	0.398	0.384	0.728	0.233
Mixtral-8x7B-v0.1	0.013	0.614	0.601	0.690	0.245
Meta-Llama-3-8B	0.013	0.547	0.534	0.660	0.255
gemini-pro	0.012	0.688	0.676	0.677	0.249
Mistral-7B-v0.1	0.011	0.431	0.420	0.583	0.280
Meta-Llama-3-70B-Instruct	0.008	0.907	0.899	0.714	0.238
Mixtral-8x22B-Instruct-v0.1	0.008	0.890	0.882	0.612	0.270
Phi-3-mini-4k-instruct	0.007	0.807	0.800	0.474	0.318
claude-3-haiku-20240307	0.006	0.792	0.785	0.416	0.339

28

Alternative Prompt

Model	Diff	GSM8k	GSM1k	Z-score	p-value
Llama-2-13b-hf	0.005	0.281	0.276	0.298	0.383
Xwin-Math-70B-V1.0	0.005	0.808	0.803	0.319	0.375
gpt2-xl	0.004	0.006	0.002	1.422	0.078
Mixtral-8x22B-v0.1	0.002	0.808	0.807	0.115	0.454
gemini-1.5-flash-preview-0514	0.001	0.810	0.808	0.110	0.456
CodeLlama-7b-Python-hf	0.001	0.119	0.118	0.112	0.455
CodeLlama-13b-Instruct-hf	-0.000	0.284	0.285	-0.028	0.511
gemma-2b-it	-0.000	0.101	0.101	-0.064	0.526
CodeLlama-34b-Instruct-hf	-0.002	0.403	0.404	-0.073	0.529
CodeLlama-13b-hf	-0.004	0.213	0.217	-0.232	0.592
Phi-3-medium-128k-instruct	-0.005	0.870	0.876	-0.368	0.644
claude-3-opus-20240229	-0.006	0.830	0.836	-0.396	0.654
claude-2.1	-0.006	0.836	0.842	-0.425	0.665
gpt-4	-0.008	0.919	0.927	-0.790	0.785
gpt-4-turbo	-0.011	0.847	0.858	-0.825	0.795
Mistral-7B-Instruct-v0.1	-0.011	0.340	0.352	-0.617	0.731
gpt-4o	-0.012	0.913	0.925	-1.188	0.882
Llama-2-70b-hf	-0.013	0.572	0.585	-0.636	0.738
gemini-1.5-pro-preview-0514	-0.014	0.802	0.816	-0.894	0.814
mistral-large-latest	-0.017	0.854	0.871	-1.228	0.890
gpt-3.5-turbo	-0.017	0.742	0.759	-0.994	0.840
claude-3-sonnet-20240229	-0.024	0.713	0.737	-1.326	0.908
Meta-Llama-3-70B	-0.034	0.815	0.849	-2.287	0.989

29

932

G 50 Examples from GSM1k

A previous version of this paper mistakenly included some questions from a nonfinal version of GSM1k. A corrected table is below.

No.	Question	Answer
1	Gabriela has $65.00 and is shopping for groceries so that her grandmother can make her favorite kale soup. She needs heavy cream, kale, cauliflower, and meat (bacon and sausage). Gabriella spends 40% of her money on the meat. She spends $5.00 less than one-third of the remaining money on heavy cream. Cauliflower costs three-fourth of the price of the heavy cream and the kale costs $2.00 less than the cauliflower. As Gabriela leaves the store, she spends one-third of her remaining money on her grandmother's favorite Girl Scout Cookies. How much money, in dollars, does Gabriela spend on Girl Scout cookies?	7
2	Bernie is a street performer who plays guitar. On average, he breaks three guitar strings a week, and each guitar string costs $3 to replace. How much does he spend on guitar strings over the course of an entire year?	468
3	John Henry is competing against a machine to see who can dig a tunnel more quickly. John works without rest, and excavates at a rate of 6 cubic feet of rock per hour. The machine excavates more quickly but needs to be refueled and maintained by its operator for 30 minutes out of every hour. When it's not under maintenance, the machine excavates at a rate of 10 cubic feet of stone per hour. Provided that the competition lasts for 8 hours, how much more rock will John have excavated compared to the machine?	8
4	Colin is playing dice with his friend Eoin and needs some help keeping track of his score. He begins with 5 points and wins 6 points in the first round. In the second round, he won twice as many points as he won in the first round. In the third round, he had a fantastic roll and was able to triple his total point count! How many points did Colin end the game with?	69
5	Marge got a job so she can buy her first car. Her job pays $15/hr and she works there 30 hours a week. The car Marge wants is $3600. How many weeks does Marge need to work to buy the car?	8
6	Andy's soccer team needs 80 points to finish in first place. His team plays 38 games, and he gets 3 points for each win, 1 point for each tie, and 0 points for each loss. After 26 games, the team has 15 wins, 5 ties, and 6 losses. How many more points does Andy's team need to reach 80 points?	30
7	Molly wants to win the contest at school for reading 25 books before the end of May. So far, she has read 5 books by the end of January. How many more books will she need to read on average each month until the end of May to win the contest?	5
8	Ms. Crabapple has a bag of jelly beans that she is going to divide equally among all of her 32 students who complete their homework every day over the course of a week. The bag has 384 jellybeans in it. Unfortunately, many of Ms. Crabapple's students have a poorly developed work ethic, and only half of them complete all of the required homework. How many jelly beans will each of the eligible students receive?	24
9	Bob has to read 2 books and 3 articles, while Emily has to read 4 books and 2 articles. Each book has 3 chapters and each chapter has 4 paragraphs. Each article has 4 sections and each section has 2 paragraphs. How many paragraphs in total will Bob and Emily read?	112

30

No.	Question	Answer
10	Leah and 2 of her friends go to an all-you-can-eat dumpling buffet. Leah's 1st friend ate 30 dumplings, her 2nd friend ate twice as many dumplings as her 1st friend, and Leah ate 1.5 times as many dumplings as her 2nd friend. How many dumplings in total did Leah and her friends eat?	180
11	Francis has a bowl of candy in front of him. There are three different flavors of candies that he's eaten over the course of 3 hours. He's eaten ten lemon, four orange, and sixteen cherry-flavored candies. If there were twenty of each when he started, how much of an average percentage is still left?	50
12	Maryann is saving up for a new bike that costs $450. She already has $120 saved up. She earns $15 per hour at her part-time job. How many hours does she need to work to afford the bike?	22
13	Henry is renovating his kitchen and adding a new tile floor. He needs to cover an area of 200 square feet. He has a stack of tiles that measure 0.5 feet in length and width. He can get 40 tiles done per hour. Henry works for 6 hours at that rate, then has some coffee and works at a faster rate for the next 2 hours (60 tiles per hour). Henry runs out of tiles, so he goes to a store to purchase the remaining tiles needed to finish the floor. Given that the price per tile is $2.50, how much will he need to spend at the store to get exactly enough tiles to finish the floor?	1100
14	A painter needs to paint 3 houses. The first house requires 14 gallons of paint, the second house requires twice as much paint as the first, and the third house needs half as much paint as the second house. If one gallon of paint costs $35 and the painter gets a bulk discount of 10% for purchases over 30 gallons, how much will the paint cost in total?	1764
15	A coal miner is loading up coal into mine carts. During the first hour of the day, he is able to load 15 carts. His boss yells at him after that, so for each of the next three hours, he loads twice as many carts. Each cart weighs 78 pounds. What was the total weight of the coal he loaded on this day?	8190
16	A plane owned by Sunny Skies Airlines is flying from Indianapolis to Phoenix. The plane holds 180 passengers and is 2/3 full. Each passenger brings 2 carry-on bags and is charged a carry-on bag fee of $35 per bag. How much money does Sunny Skies Airlines collect for the carry-on bag fees for this flight?	8400
17	Sally went to the mall to buy clothes for the summer. She went to Forever 21 and bought 4 tops, each had different prices, $12.99, $6.99, $17.99, $21.99, and 3 pants each priced at $15.99. If her subtotal is over $75, she gets a discount of 15% on her purchase at that store. Then she goes to Shoe Palace and buys 2 shoes for a total of $123.26. How much money did Sally spend at the mall?	215
18	Dean wants to buy flowers to make arrangements for a party. He is going to make 12 arrangements. He wants to include 4 roses and 3 daisies in each arrangement. Roses come by the dozens and are $15 for each dozen. Daisies come in groups of 4 and are $8 for the set. How much will it cost for Dean to make all 12 arrangements?	132

31

No.	Question	Answer
19	Alex plans to adopt a new cat and needs help planning a budget for this event. The adoption fee is $200, and it includes all the essential veterinary care needed for a kitten, but she also needs to buy other supplies for the cat when she brings it home. The litter boxes cost $30, one package of litter costs $17, a bag of dry food costs $55, and the wet food costs $1.50 per can. Alex will buy 2 litter boxes, 3 packages of litter, one bag of dry food, and 12 cans of wet food. How much money should Alex make sure she has before beginning the process of adopting her new cat?	384
20	Samantha is saving money for a new bike by doing chores. She earns $5 for every chore she completes. If she does 3 chores each day for a week, and then uses $25 to buy a helmet, how much money does she have left at the end of the week?	80
21	Frank sneaks out before his break at 3:20 pm and gets back at 4:05. If his break was only supposed to be half an hour, for how much longer did Frank sneak out?	15
22	Janet wants to listen to 20 music albums by the end of the week. If she just finished her twelfth album and today is Thursday, how many albums per day would she have to listen to by Saturday?	4
23	Hana wants to donate her clothes to a local charity. After going through her closet she ended up with 2 boxes of pants, 3 boxes of dresses, 1 box of shoes, and boxes of shirts. The number of boxes with shirts was 3 more than the other three boxes combined. How many boxes of shirts does she have to donate?	9
24	Gayle has a lawnmowing business. Lawn 1 takes 15 minutes to mow. Lawn 2 takes 18 more minutes than Lawn 1. Lawn 3 takes 20% more time to mow than Lawn 1. She is paid $2.50 per minute for the time she spends. However, she gives her customers a 20% discount. How much money does she make from mowing all three lawns?	132
25	Frank ordered a whole chicken, 6 cans of chopped chicken breast, 1 lb. of macadamia nuts, and 4 bags of frozen broccoli. Each item has the following respective prices: $12 per chicken, $2 per can, $24/lb., $3 per bag. The sales tax was 10% of the total cost and the tip was half the price of the whole chicken. How much did Frank pay for his order?	72
26	Milo can bench press half as much weight as Doug can squat, and Doug can squat twice as much weight as Diane can squat. If Diana squats 125 pounds, how much weight can Milo bench press?	125
27	Pablo is trying to make breakfast for his family. His wife eats 4 pancakes. His son eats 2 pancakes. Pablo wants to eat 4 pancakes. One box of pancake mix will make 5 pancakes. How many boxes of pancake mix will he need?	2
28	Jim wants to spend 15% of his monthly earnings on groceries. He makes $2500/month. How much money will he have left over?	2125
29	A school is ordering tablets and laptops for three classrooms. Each classroom will receive 4 tablets and 3 laptops. If each tablet costs $250 and each laptop costs $600, how much will the school spend in total for all three classrooms?	8400
30	Grant takes 3 minutes to put on his pajamas. He brushes his teeth for 2 minutes. Then, he washes his face and brushes his hair for another 2 minutes. Finally, he reads a book for a while and turns off the light for bed. If Grant begins his routine at 8:15 pm and turns off the lights at 8:47 pm, for how long does Grant read a book?	25

32

No.	Question	Answer
31	Bellemere owns a tangerine orchard with 50 trees. Each tree produces 80 tangerines. She wants to sell 600 tangerines at her local farmer's market. If she picks the same amount of tangerines from every tree, how many tangerines will be left on each tree?	68
32	A charity puts out a telethon for a cause. Within 15 minutes, seventy-seven people donated $3 each, and 231 people donated four dollars each. How much does the charity receive within this time?	1155
33	A school is selling baskets for a fundraiser. There are three baskets containing the following items: * Blue basket: a ball, cup, and notebook. * Red basket: a cup, bell, and hat. * Green basket: a hat, pen, and notebook. The costs of the items in the baskets are as follows: * $1: ball, notebook, and pen * $2: cup, bell, and hat Jane buys 6 red baskets and 5 blue baskets. Jim buys 3 red baskets and 2 green baskets. Since they purchase so many, they receive a discount. Jane gets an $8 discount and Jim also gets a $2 discount. How many times more does Jane spend than Jim?	2
34	Mr. Gordon has 14 boys in his first period class which is twice the number of girls in class. Two of the girls in class have blonde hair and the rest have brown hair. How many girls with brown hair are in his class?	5
35	Albert gets paid $15 an hour. He gets time and a half if he works over forty hours a week. Last week, he worked 48 hours. He plans to do this two weeks in a row. How much money will he be paid in overtime for those two weeks?	360
36	Beth, Anna, and Kim went to a book fair. Beth had two books less than Anna while Kim had four more books than Anna. Beth had $20 with her and was now left with $8. If all books are priced at $4, how much, in dollars, did Kim spend on her books?	36
37	4 friends are going on a road trip. Their names are Alex, Bethany, Carlos, and Drew. They drive at a rate of 65, 75, 60, and 50 mph, respectively. Alex drives for 2 hours, Bethany for 4, and Carlos and Drew each drive for 3 hours. They are using a car with a fuel efficiency of 20 miles per gallon of gas. If, along their route, gas costs $3 per gallon, how much money (in dollars) will they need to spend on gas? Assume they begin their journey at a gas station with an empty tank of gas.	114
38	The Genco Olive Oil Company has received ninety-nine orders for ninety-nine barrels of olive oil each. Out of those shipped, 33 orders were sent back due to clerical or product errors. How many total barrels of olive oil were not returned?	6534
39	There is a very large room that has 4 tables, 1 sofa and 2 chairs that have 4 legs each. There are also 3 tables with 3 legs each, 1 table with 1 leg, and 1 rocking chair with 2 legs. How many legs of tables are there in the room?	26
40	A classroom has 24 students, and the teacher has arranged a field trip. If the cost per student for the trip is $15 and the teacher already has $120 from a class fund, how many more dollars does the teacher need to cover the total cost of the trip for all students?	240
41	Rachel and Shauna go out to dinner. Dinner costs $68.25 in total (without taxes). Rachel's meal costs 1/3 of the total price, while Shauna's meal costs 2/3 of the total price. How much did Shauna's meal cost (round to the nearest dollar)?	46

33

No.	Question	Answer
42	Olivia owns a local hotel and needs to drive up business. She is planning to give a special deal to anyone who signs up for a membership card. Her idea is to give them 20% off their first night and 10% off on every night they stay after that. If her first new customer pays $616 for their stay, and each night costs $140 before discounts, how many nights did they stay at the hotel?	5
43	Johnny has 8 green balls. He has five fewer than twice that in red balls. How many total balls does Johnny have?	19
44	30 students are in a class. 1/5 of them are 12 years old, 1/3 are 13 years old. 1/10 of them are 11 years old. How many of them are not 11, 12, or 13 years old?	11
45	Francis loves sandwiches. He gets his usual from his favorite deli: two "Big Boy" sandwiches, and a glass-bottled soda. A "Big Boy" costs $15.25 and the soda costs $3.75. His friend Lars calls him and asks for a double-sweet soda that's $4.75. If Francis pays all of this with $40 and asks for his change back in only quarters, how many quarters will he get?	4
46	A factory needs to produce 960 pieces of toy boats. They are only able to produce 1/6th of their goal a day. 5 toy boats make up a case and 4 cases make up a box. If a toy shop comes to pick up what is available on the fourth day and finds an extra 8 boxes left for them that were forgotten from a previous pickup, how many boxes of toy boats will they be able to take?	40
47	The highest temperature ever recorded on Earth was 136 degrees Fahrenheit and the coldest temperature ever measured was -126 degrees Fahrenheit. If the average temperature of Earth is 59, what would be the difference between the average temperature on Earth and the average given the two extremes?	54
48	Maria was shopping for the perfect prom dress. She found a red one that cost $250 but was on sale for 20% off. Sales tax is 5%. Her grandmother gave her $300 to pay for her dress and dinner. After Maria purchased the red dress, how much did she have left, in dollars, to pay for dinner?	90
49	Mrs. Watson, a high school Spanish teacher, is required to input 2 grades a week per student per her school's grading policy. Mrs. Watson has 6 classes in total. Her 1st-period class has 32 students, 2nd-period has 28 students, 3rd-period has 41 students, 4th-period has 23 students, 5th-period has 18 students, and her 6th-period class has 33 students. How many grades does Mrs. Watson need to input each week to remain compliant with the school's grading policy?	350
50	Mary sells 4 bags of pears where each bag contains 3 giant pears and 4 small pears. Each giant pear is sold at $5 and each small pear is sold at $2. Mary also sells 2 bags of cherries where each bag contains 5 pounds of cherries. The cherries are sold at $8 per pound. How much in total does Mary earn from selling all these fruits?	172

34

935
936
H Bar Chart of Performance Gaps Between GSM8k and GSM1k Across All Model Accuracies
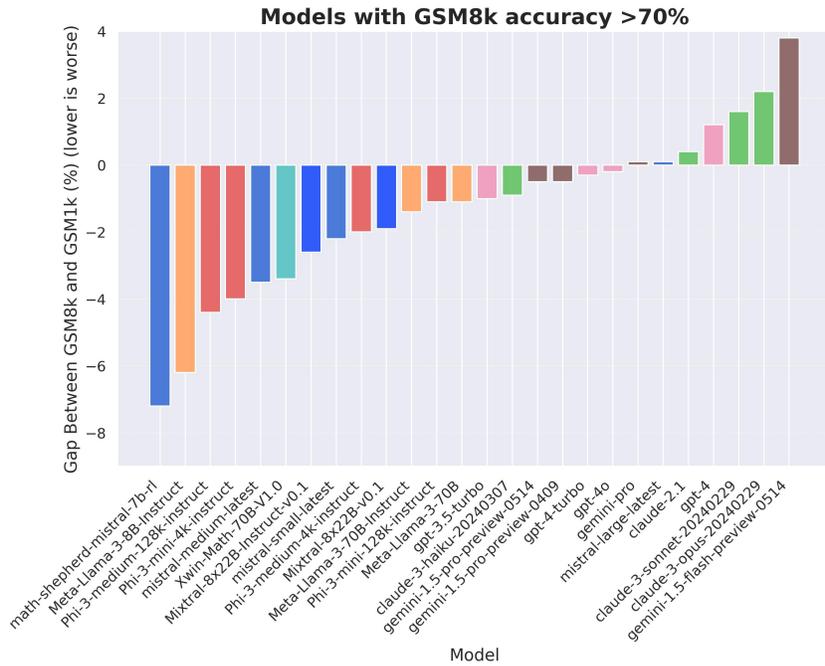
Figure 8: Models with over 70% accuracy on GSM8k. We observe that some models (e.g. Mistral, Phi) are overfit, while other models show little to no evidence of overfitting.
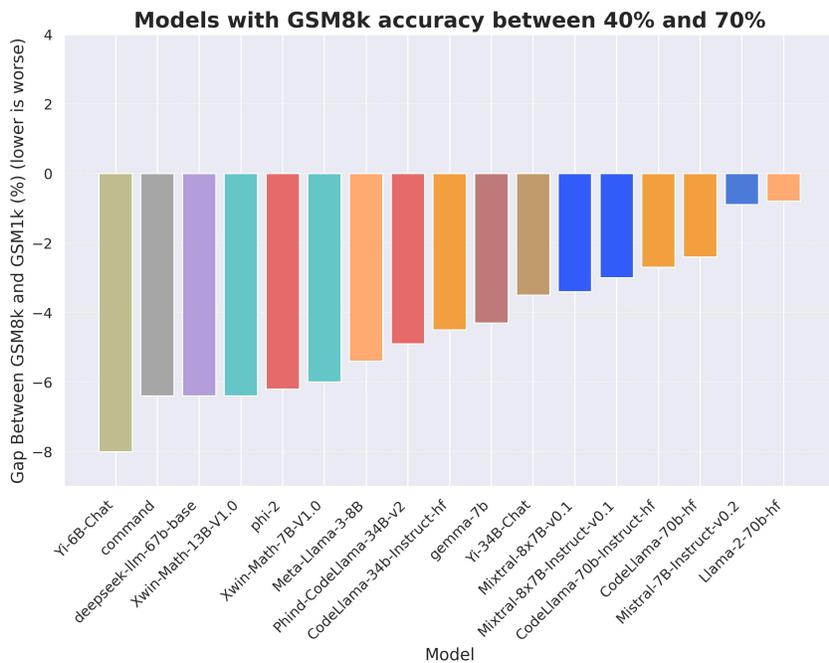
Figure 9: Comparison of models with between 40 and 70% accuracy on GSM8k. We observe that all models seem to fall below the line in this regime of model performance, though some models (e.g. Llama-2-70b) do much better than others.

35
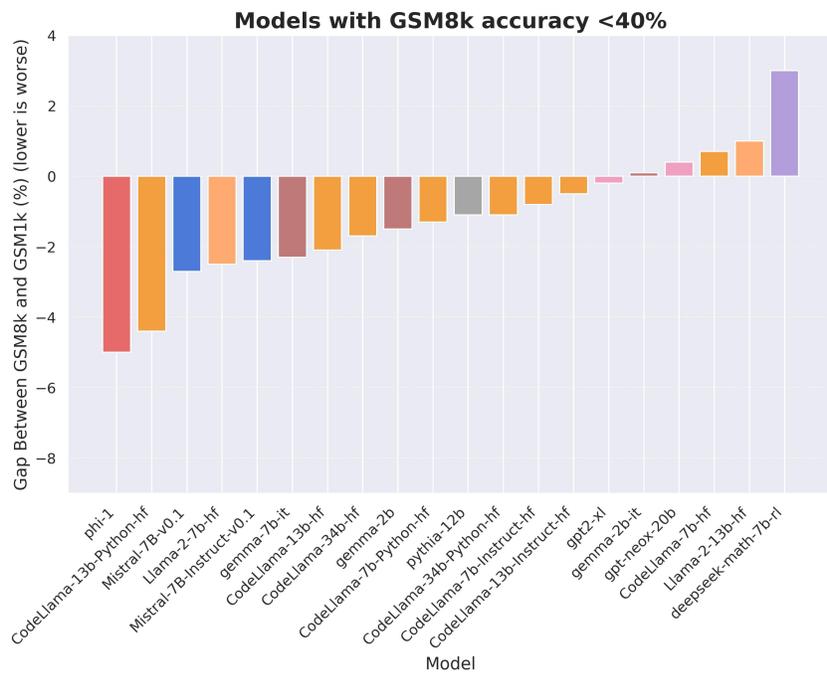
Figure 10: Models with less than 40% accuracy on GSM8k.

36

937 # I   Additional Plots

Models with GSM8k accuracy between 40% and 70%
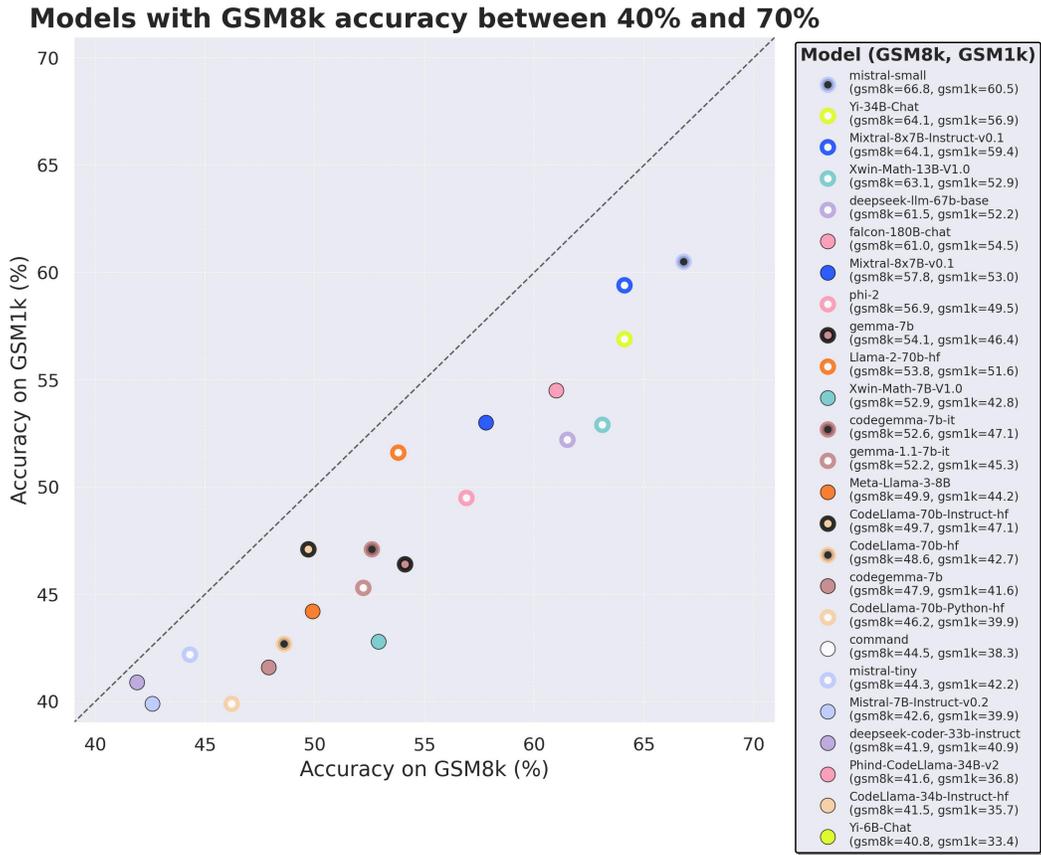
Figure 11: Models with between 40 and 70% accuracy on GSM8k compared to the line of no overfit. This plot is zoomed into the relevant sections (40-70% accuracy). We observe that no models lie on the line of no overfit in this regime.

37
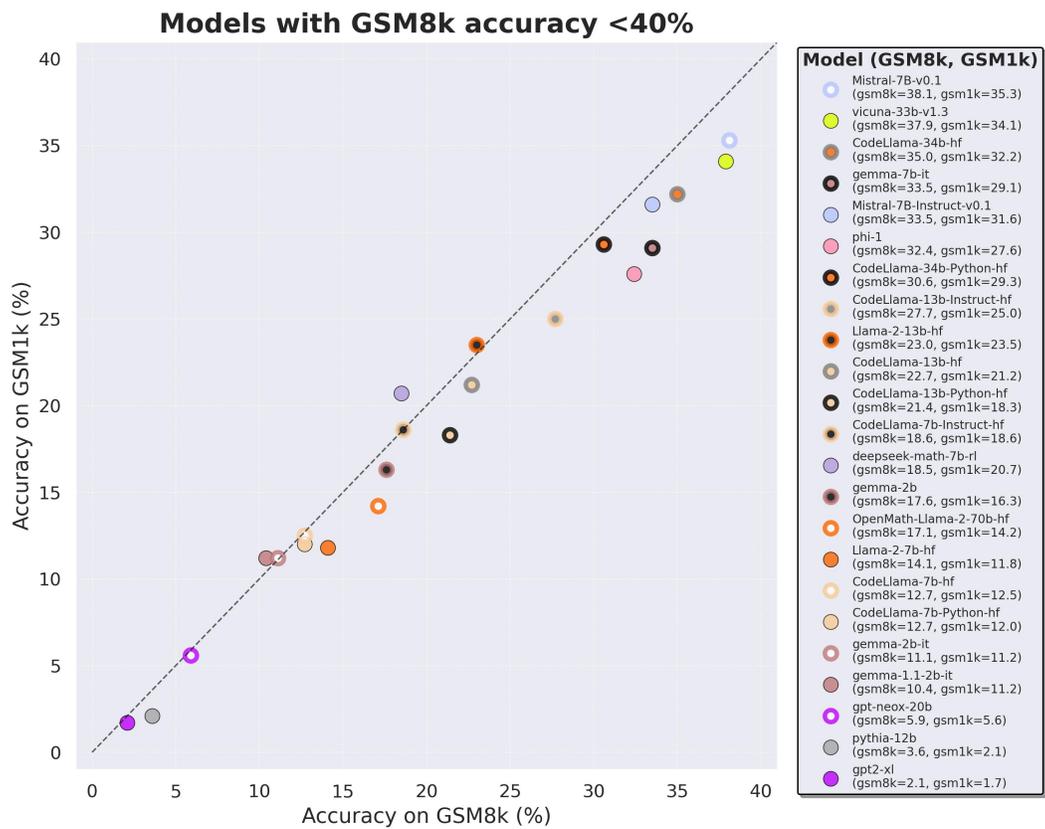
Figure 12: Models with between 0 and 40% accuracy on GSM8k compared to the line of no overfit. This plot is zoomed into the relevant sections (0-40% accuracy).

38
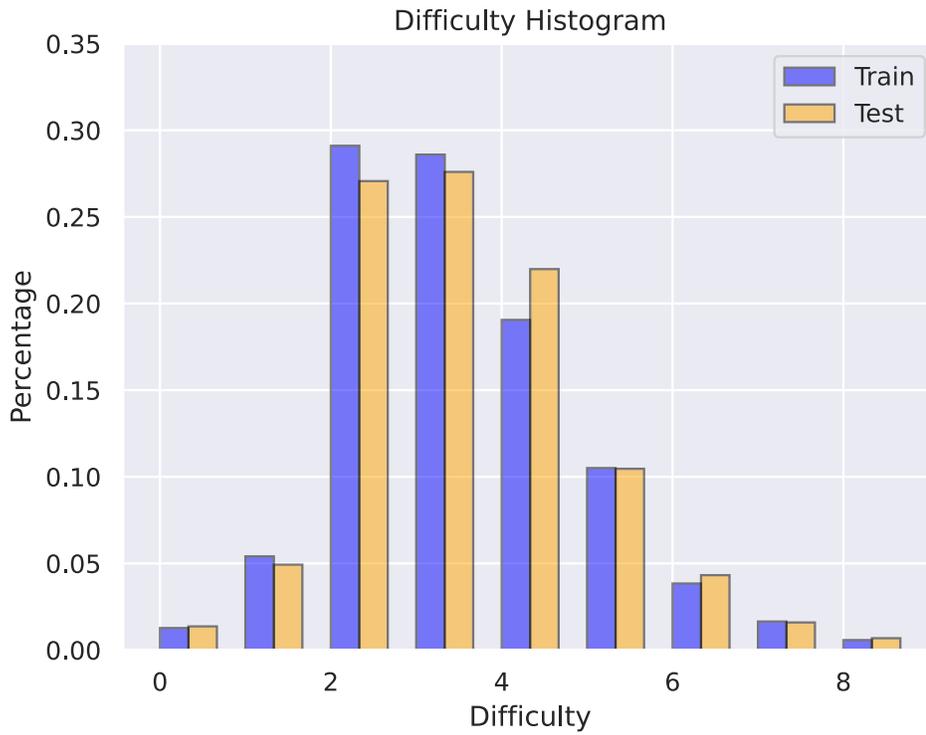
Figure 13: Approximate difficulty distribution of GSM8k train and test sets, measured by number of required steps to solve the problem. GSM1k annotators were instructed to create problems matching the overall distribution of the combined train and test difficulty distribution. The process of estimating problem difficulty is described in Section 3.2.

39

J   Ablation with Human Answer Extraction

Because LM Evaluation Harness uses an automatic extraction format which takes the last number outputted as the model's final answer and compares an extract string match with the gold standard. Models answers which fail to follow the proper format may be marked as incorrect even if the model produced the "correct" answer. This most frequently occurs when the model phrases its answer in natural language, but adds extraneous information at the end of its sentence. Additionally, automatic extraction is brittle: in a few cases we observe that model outputs such as "24.0" are marked as not matching the gold standard answer of 24 due to the trailing decimal points failing the exact string match.

To measure the impact of extraction errors, we analyze a subset of models and use human annotators to extract model answers. These models were chosen for high performance on GSM1k for the purposes of creating a leaderboard of top performing models. While the absolute performance numbers change, we do not find meaningful differences in the amount of overfitting between GSM1k and GSM8k based on whether human or automatic extraction was used for this set of models.

Model	GSM1k (Human)	GSM1k (Auto)	GSM8k (Human)	GSM8k (Auto)	Gap (Human)	Gap (Auto)
Claude 3 Opus	0.952	0.825	0.955	0.802	-0.003	0.023
GPT-4 Turbo Preview	0.951	0.898	0.952	0.898	-0.001	0.0
GPT-4o	0.949	0.928	0.962	0.933	-0.014	-0.005
Claude 3 Sonnet	0.933	0.744	0.926	0.719	0.006	0.024
Gemini 1.5 Pro (post-I/O)	0.923	0.895	0.933	0.915	-0.01	-0.02
Gemini 1.5 Pro (pre-I/O)	0.905	0.885	0.904	0.897	0.002	-0.011
Llama 3 70B Instruct	0.901	0.895	0.917	0.896	-0.016	-0.001
Gemini 1.5 Flash	0.901	0.832	0.896	0.804	0.005	0.027
Mistral Large	0.875	0.853	0.892	0.853	-0.017	0.0
Gemini 1.0 Pro	0.798	0.789	0.805	0.792	-0.007	-0.002
CodeLlama 34B Instruct	0.375	0.366	0.422	0.415	-0.047	-0.049

40

K Ablations with the Alternative Format

We investigate the impact of prompting on several of the model families with highest amounts of overfitting. In this section, we test whether the difference in performance with the standard and alternative prompt is due to the alternative prompt using non-GSM8k examples. We do this by constructing prompts in the same format as the alternative "chain-of-thought" prompt but using fewshot example problems randomly chosen from GSM8k. These prompts imitate the alternative prompt's answer format and use of 8 fewshot examples rather than 5 in the standard prompt. We do this by converting two sets of randomly selected GSM8k problems into the analogous format.

We find significant variance in the results, ablating even something as simple as *which* n-shot examples are chosen. Nevertheless, the general shape of the findings remains largely consistent, even if the precise ordering / numerical values are highly prompt dependent.

For the first such ablation prompt (provided in full below), our results are displayed in Figure 14

```
Q: Bob drove for one and a half hours at 60/mph.  He then hit
   construction and drove for 2 hours at 45/mph. How many miles did Bob
   travel in those 3 and a half hours?
A: Bob drove for 1.5 x 60 = 90 miles first, then another 2 x 45 = 90
   miles. In total Bob drove 90 + 90 = 180 miles. The answer is 180.

Q: Mary is paying her monthly garbage bill for a month with exactly four
   weeks. The garbage company charges Mary $10 per trash bin and $5 per
   recycling bin every week, and Mary has 2 trash bins and 1 recycling
   bin. They're giving her an 18% discount on the whole bill before
   fines for being elderly, but also charging her a $20 fine for putting
    inappropriate items in a recycling bin. How much is Mary's garbage
   bill?
A: Every week, Mary pays 10 x 2 = 20 dollars for the trash bins, so her
   total weekly cost is 20 + 5 = 25 dollars for the trash bins and the
   recycling bin. Then her monthly cost over 4 weeks is 25 x 4 = 100
   dollars. Mary's senior discount is 18 x .01 x 100 = 18 dollars. So
   subtracting the discount and adding the fine, her total monthly cost
   is 100 − 18 + 20 = 102 dollars. The answer is 102.

Q: June has $500 for buying school supplies for the new school year. She
   buys four maths books at $20 each, six more science books than maths
   books at $10 each, and twice as many art books as maths books at $20
   each. If she also bought music books, how much money did she spend on
    music books?
A: The total cost of maths books is 4 x 20 = 80. She bought six more
   science books than maths books which totals 6 + 4 = 10 books. If each
    science book cost her $10 she spent 10 x 10 = 100 dollars on science
    books. There were twice as many art books as maths books which total
    2 x 4 = 8. The total cost for art books is 8 x 20 = 160. The total
   amount that she used for maths, science, and art books is 160 + 100 +
    80 = 340. The amount she spent on music books is 500 − 340 = 160.
   The answer is 160.

Q: A play was held in an auditorium and its ticket costs $10. An
   auditorium has 20 rows and each row has 10 seats. If only 3/4 of the
   seats were sold, how much was earned from the play?
A: There are 20 x 10 = 200 seats in the auditorium. Only 200 x 3/4 = 150
   seats were sold. Hence, the earnings from the play is 10 x 150 =
   1500. The answer is 1500.

Q: Brendan went fishing with his dad. Brenden caught 8 fish in the
   morning. He threw 3 back that were too small. He caught 5 more in the
    afternoon. Brendan's dad caught 13 fish. How many fish did they
   catch in all?
A: Brenden caught 8 fish in the morning and 5 in the afternoon so he
   caught 8 + 5 = 13 total fish. After throwing the small fish back,
```

41

```
1011    Brenden has 13 − 3 = 10 fish. Together, Brenden and his dad caught 10
1012     + 13 = 23 fish. The answer is 23.
1013
1014 Q: Valerie's cookie recipe makes 16 dozen cookies and calls for 4 pounds
1015    of butter.  She only wants to make 4 dozen cookies for the weekend.
1016    How many pounds of butter will she need?
1017 A: Her original recipe makes 16 dozen and she only needs 4 dozen so she
1018    needs to reduce the recipe by 16 / 4 = 4. For 4 dozen cookies, she
1019    needs to reduce her recipe by 4 and the original called for 4 pounds
1020    of butter so she now needs 4 / 4 = 1 pound of butter. The answer is
1021    1.
1022
1023 Q: Jack is mad at his neighbors for blasting Taylor Swift all night, so
1024    he slashes three of their tires and smashes their front window. If
1025    the tires cost $250 each and the window costs $700, how much will
1026    Jack have to pay for the damages?
1027 A: The total cost of the tires is 250 x 3 = 750. Then total cost of the
1028    tires and the window is 700 + 750 = 1450. The answer is 1450.
1029
1030 Q: A dental office gives away 2 toothbrushes to every patient who visits.
1031     His 8 hour days are packed and each visit takes .5 hours.  How many
1032    toothbrushes does he give in a 5 day work week?
1033 A: Each day he does 8 / .5 = 16 visits. So he does 16 x 5 = 80 visits a
1034    week. That means he gives away 80 x 2 = 160 toothbrushes a week. The
1035    answer is 160.
```
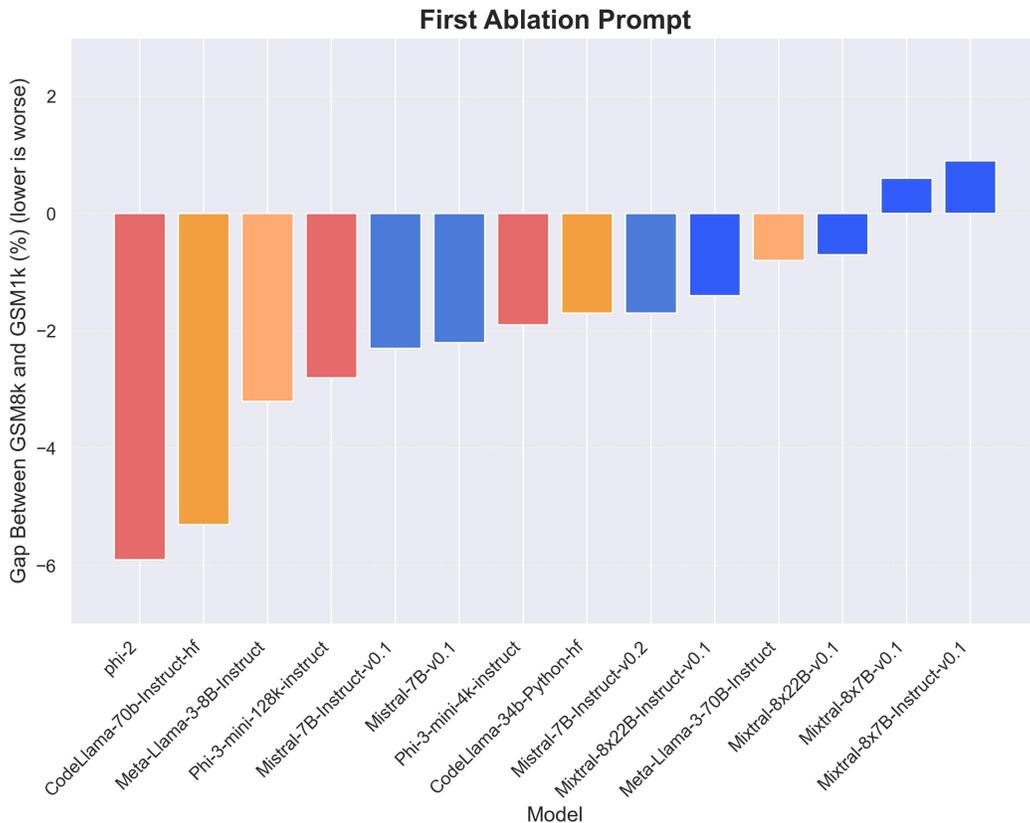
Figure 14: Models from the most overfit families arranged by their drop in performance between GSM8k and GSM1k (lower is worse) on the first ablation prompt with GSM8k examples in the alternative format.

1036   For the second such ablation prompt (provided in full below), our results are displayed in Figure 15

42

Q: At camp Wonka, there are 96 campers. Two—thirds of the campers are
   boys, and the remaining one—third are girls. 50% of the boys want to
   toast marshmallows and 75% of the girls want to toast marshmallows.
   If each camper gets one marshmallow to toast, how many marshmallows
   do they need?
A: The girls make up one—third of the campers, so there are 96 / 3 = 32
   girls. The boys make up two—thirds of the campers, so there are 32 +
   32 = 64 boys. There are 32 x 75% = 24 girls who want to toast
   marshmallows. There are 64 x 50% = 32 boys who want to toast
   marshmallows. They need 24 + 32 = 56 marshmallows. The answer is 56.

Q: In today's field day challenge, the 4th graders were competing against
   the 5th graders. Each grade had 2 different classes. The first 4th
   grade class had 12 girls and 13 boys. The second 4th grade class
   had 15 girls and 11 boys. The first 5th grade class had 9 girls and
   13 boys while the second 5th grade class had 10 girls and 11 boys.
   In total, how many more boys were competing than girls?
A: When you add up all the girls from all 4 classes, you had 12 + 15 + 9
   + 10 = 46 girls. When you add up all the boys from all 4 classes, you
   had 13 + 11 + 13 + 11 = 48 boys. There are 48 boys and 36 girls so
   48 − 46 = 2 more boys. The answer is 2.

Q: Axel bought an aquarium that was marked down 50% from an original
   price of $120. But he also paid additional sales tax equal to 5% of
   the reduced price. What was the total cost of the aquarium?
A: The aquarium was bought for $120 x 50/100 = $60 less. So the marked
   down price of the aquarium was $120 − $60 = $60. Axel paid $60 x
   5/100 = $3 additional for the sales tax. Therefore, the total cost of
   the aquarium was $60 + $3 = $63. The answer is 63.

Q: There are 48 crayons in the box. Kiley takes 1/4 of them away. Joe
   takes away half of the remaining crayons, how many crayons are left?
A: Kiley takes 48 / 4 = 12 crayons, so 48 − 12 = 36 crayons remain. Joe
   takes 36 / 2 = 18 crayons, so there are 36 − 18 = 18 crayons left.
   The answer is 18.

Q: Six Grade 4 sections launched a recycling drive where they collect old
   newspapers to recycle. Each section collected 280 kilos in two weeks
   . After the third week, they found that they need 320 kilos more to
   reach their target. How many kilos of the newspaper is their target?
A: In a week, each section collected 280 / 2 = 140 kilos of newspapers.
   So, in three weeks, one section collected 140 x 3 = 420 kilos. So,
   the four sections collected a total of 420 x 4 = 1680 kilos. Hence,
   their target is to collect 1680 + 320 = 2000 kilos of the newspaper.
   The answer is 2000.

Q: Jeff has a shelter where he currently takes care of 20 cats. On Monday
   he found 2 kittens in a box and took them to the shelter. On Tuesday
   he found 1 more cat with a leg injury. On Wednesday 3 people adopted
   2 cats each. How many cats does Jeff currently have in his shelter?
A: Counting the cats he had, the kittens he found, and the injured cat,
   Jeff had a total of 20 + 2 + 1 = 23 cats. 3 people took a total of 3
   x 2 = 6 cats. After Wednesday, Jeff was left with 23 − 6 = 17 cats.
   The answer is 17.

Q: Paul is working at a university. He is part of a big project, which
   involves 70 scientists in total. Half of them are from Europe and one
   —fifth are from Canada. The rest are from the USA. How many
   scientists in this project are from the USA?
A: Of all the scientists taking part in the project, half of them are
   from Europe, which means 70 x 0.5 = 35 people. The number of
   researchers from Canada is 70 x 1/5 = 14 people. That means there are
   70 − 35 − 14 = 21 researchers from the USA. The answer is 21.

43

```
1101  Q: John buys a heating pad for $30.  He uses it 3 times a week for 2
1102     weeks.  How much does he spend on each use?
1103  A: He uses it 3 x 2 = 6 times. So he pays $30 / 6 = $5 for each use. The
1104     answer is 5.
```
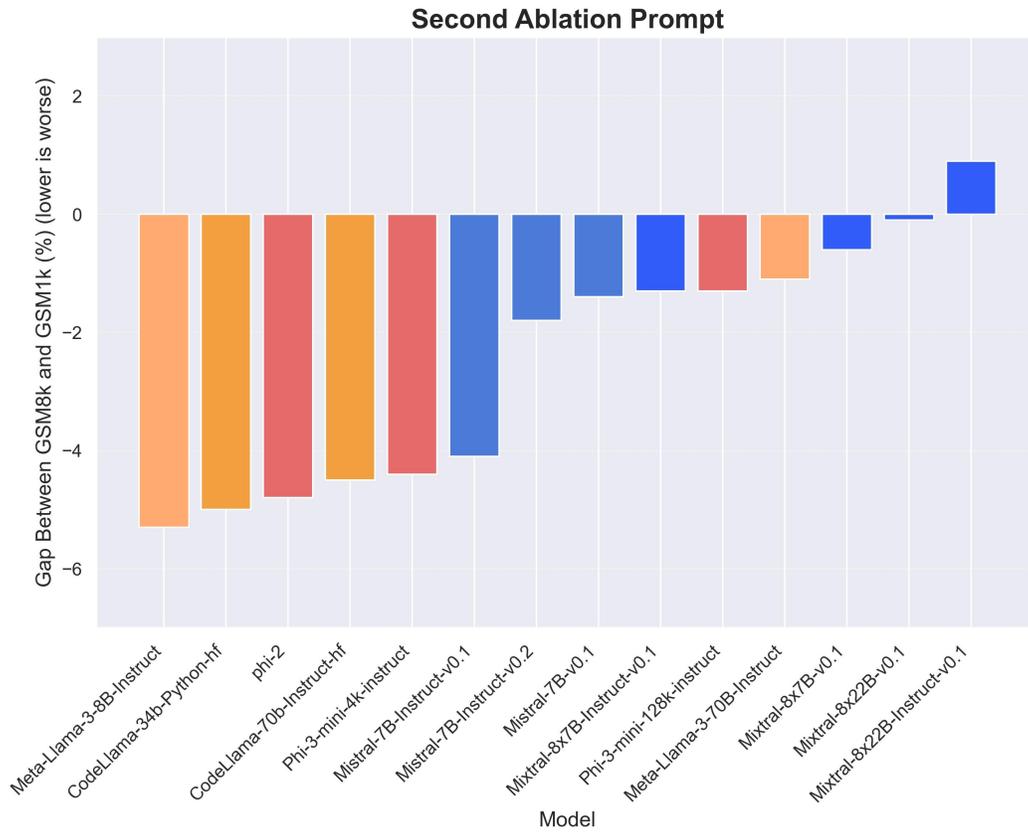
Figure 15: Models from the most overfit families arranged by their drop in performance between GSM8k and GSM1k (lower is worse) on the second ablation prompt with GSM8k examples in the alternative format.

44

L   Ablations with Number of Fewshot Examples

As another ablation, we also investigate the impact of varying the number of GSM8k fewshot examples in the standard prompt format. We evaluate the models from the most overfit model families on the standard n-shot prompt, with n varying from 1 to 10, inclusive. As in the primary results, the n fewshot examples in the prompt are randomly selected from GSM8k train, and vary among the questions from the GSM1k/GSM8k test set. Note that the primary findings correspond to n=5. Our results are displayed in Figure 16. We notice that all models display a performance gap between GSM8k and GSM1k for almost all values of n, demonstrating a robustness to the overfitting result despite the variance from using different prompts.
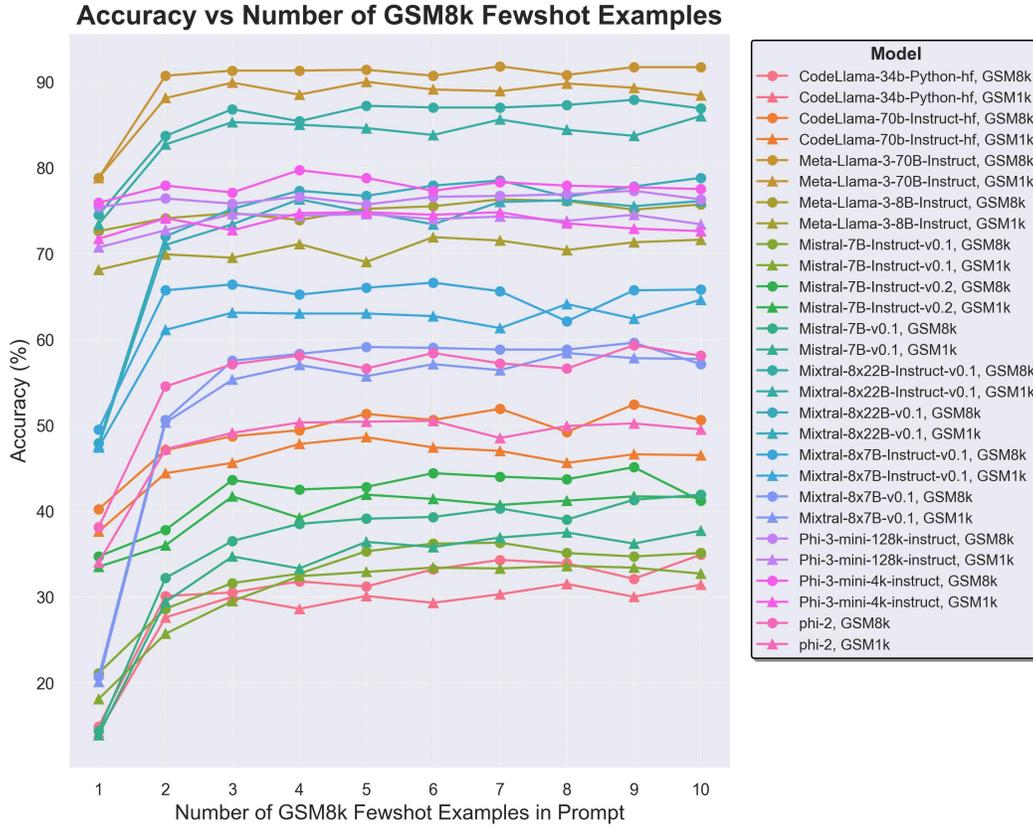
Figure 16: Performance on GSM8k and GSM1k relative to the number of GSM8k fewshot examples given in the standard prompt format, for models from the most overfit model families.

M List of Models Evaluated Using Default HuggingFace

We used vLLM to speed up model inference. A small number of models were not supported by vLLM at the time of initial evaluation, so we used the regular HuggingFace libraries to generate results. The list of these models is below.

- databricks/dbrx-base
- databricks/dbrx-instruct
- google/gemma-2b
- google/gemma-7b
- google/gemma-7b-it
- google/gemma-2b-it
- google/gemma-1.1-7b-it
- google/gemma-1.1-2b-it
- google/codegemma-7b
- google/codegemma-7b-it
- microsoft/Phi-3-mini-4k-instruct
- microsoft/Phi-3-mini-128k-instruct
- microsoft/Phi-3-medium-4k-instruct
- microsoft/Phi-3-medium-128k-instruct

46