NEURAL INFORMATION
PROCESSING SYSTEMS

OpenMathInstruct-1: A 1.8 Million Math Instruction Tuning Dataset

Shubham Toshniwal

Ivan Moshkov

Sean Narenthiran

Daria Gitman

Fei Jia

Igor Gitman

NVIDIA

Math Reasoning Task

Problem

Aaron has four brothers. If the number of Bennett’s brothers is two less than twice the number of Aaron’s brothers, what is the number of Bennett’s brothers?

Math Reasoning Task

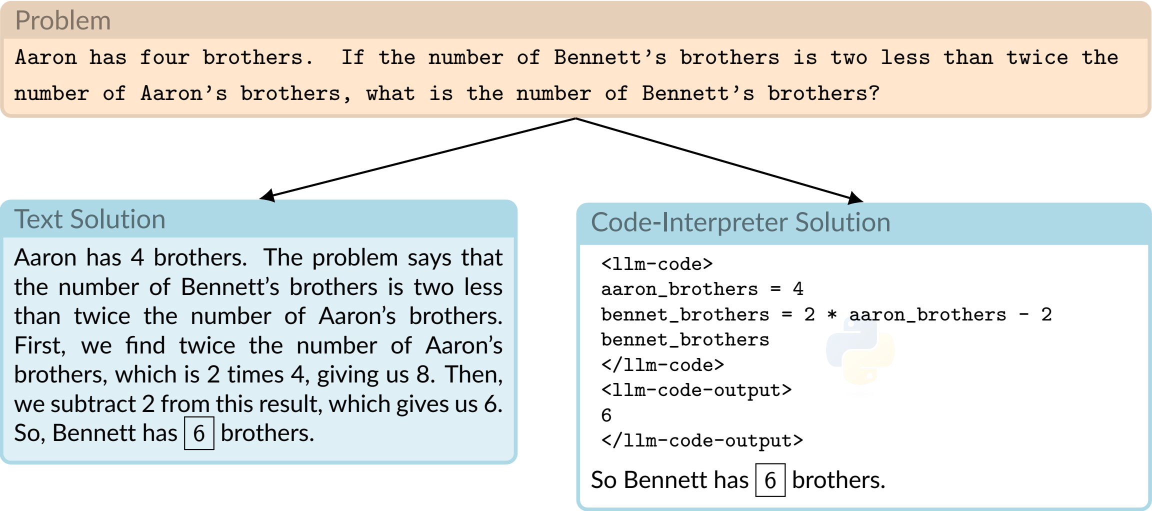
Problem

`Aaron has four brothers. If the number of Bennett's brothers is two less than twice the number of Aaron's brothers, what is the number of Bennett's brothers?`

Text Solution

Aaron has 4 brothers. The problem says that the number of Bennett's brothers is two less than twice the number of Aaron's brothers. First, we find twice the number of Aaron's brothers, which is 2 times 4, giving us 8. Then, we subtract 2 from this result, which gives us 6. So, Bennett has $\boxed{6}$ brothers.

Math Reasoning Task



Problem

```
Aaron has four brothers.  If the number of Bennett's brothers is two less than twice the number of Aaron's brothers, what is the number of Bennett's brothers?
```

Text Solution

Aaron has 4 brothers. The problem says that the number of Bennett's brothers is two less than twice the number of Aaron's brothers. First, we find twice the number of Aaron's brothers, which is 2 times 4, giving us 8. Then, we subtract 2 from this result, which gives us 6. So, Bennett has $\boxed{6}$ brothers.

Code-Interpreter Solution

```
<llm-code>
aaron_brothers = 4
bennet_brothers = 2 * aaron_brothers - 2
bennet_brothers
</llm-code>
<llm-code-output>
6
</llm-code-output>
```

So Bennett has $\boxed{6}$ brothers.

Few-shot performance on MATH: Open-Source vs Closed-Source

Model	Accuracy (%) on MATH
LLAMA-2 70B	13.8
Mistral 7B	12.7
Mixtral 8x7B	**28.4**
GPT-4	53.9
GPT-4 + Code	**69.7**

Few-shot performance on MATH: Open-Source vs Closed-Source

Model	**Accuracy (%) on MATH**
LLAMA-2 70B	13.8
Mistral 7B	12.7
Mixtral 8x7B	**28.4**
GPT-4	53.9
GPT-4 + Code	**69.7**

40 point gap between the SOTA open-source model and GPT-4 in Feb 2024

Synthetic Data to the Rescue

To bridge the gap between the open-source models and closed-source models:

- Sample solutions for the training set problems of benchmark datasets by few-shot prompting a *teacher* LLM
- Filter solutions that lead to ground truth answers
- Finetune an open-source LLM on the filtered dataset

Synthetic Data to the Rescue

To bridge the gap between the open-source models and closed-source models:

- Sample solutions for the training set problems of benchmark datasets by few-shot prompting a *teacher* LLM

- Filter solutions that lead to ground truth answers

- Finetune an open-source LLM on the filtered dataset

Best open-source models are ALL *gpt-distilled*, i.e., fine-tuned on solutions generated by GPT-4 - MetaMath (Yu et al. 2024); MAmmoTH (Yue et al. 2024)

Limitations of GPT-Distillation

Legal restraints - Distilled models can’t compete against OpenAI

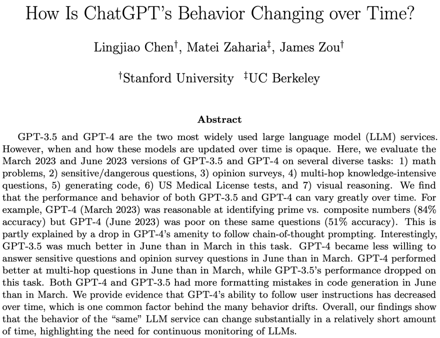
Cost - Inference with GPT-4 can cost much higher than open-source alternatives

Limitations of GPT-Distillation

Legal restraints - Distilled models can’t compete against OpenAI

Cost - Inference with GPT-4 can cost much higher than open-source alternatives

Lack of reproducibility - API behavior may change or become unavailable over time



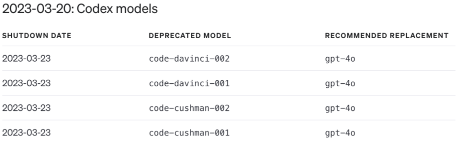
How Is ChatGPT’s Behavior Changing over Time?

Lingjiao Chen[†], Matei Zaharia[‡], James Zou[†]

[†]Stanford University [‡]UC Berkeley

Abstract

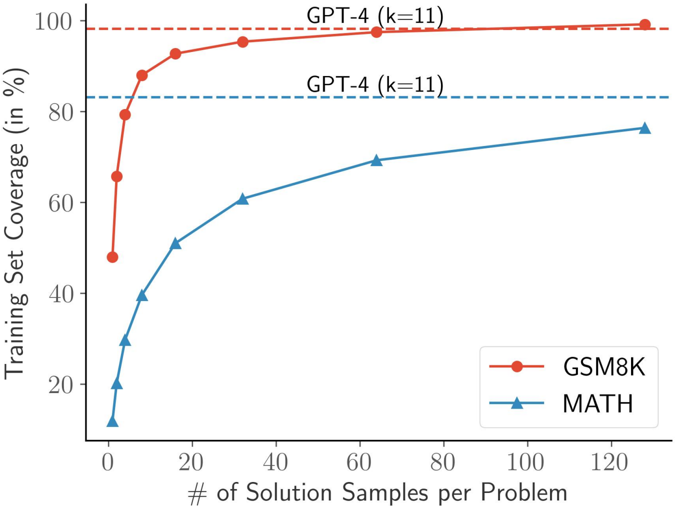
GPT-3.5 and GPT-4 are the two most widely used large language model (LLM) services. However, when and how these models are updated over time is opaque. Here, we evaluate the March 2023 and June 2023 versions of GPT-3.5 and GPT-4 on several diverse tasks: 1) math problems, 2) sensitive/dangerous questions, 3) opinion surveys, 4) multi-hop knowledge-intensive questions, 5) generating code, 6) US Medical License tests, and 7) visual reasoning. We find that the performance and behavior of both GPT-3.5 and GPT-4 can vary greatly over time. For example, GPT-4 (March 2023) was reasonable at identifying prime vs. composite numbers (84% accuracy) but GPT-4 (June 2023) was poor on these same questions (51% accuracy). This is partly explained by a drop in GPT-4’s amenity to follow chain-of-thought prompting. Interestingly, GPT-3.5 was much better in June than in March in this task. GPT-4 became less willing to answer sensitive questions and opinion survey questions in June than in March. GPT-4 performed better at multi-hop questions in June than in March, while GPT-3.5’s performance dropped on this task. Both GPT-4 and GPT-3.5 had more formatting mistakes in code generation in June than in March. We provide evidence that GPT-4’s ability to follow user instructions has decreased over time, which is one common factor behind the many behavior drifts. Overall, our findings show that the behavior of the “same” LLM service can change substantially in a relatively short amount of time, highlighting the need for continuous monitoring of LLMs.



2023-03-20: Codex models

SHUTDOWN DATE	DEPRECATED MODEL	RECOMMENDED REPLACEMENT
2023-03-23	code-davinci-002	gpt-4o
2023-03-23	code-davinci-001	gpt-4o
2023-03-23	code-cushman-002	gpt-4o
2023-03-23	code-cushman-001	gpt-4o

Why not use Open-Source Models? Mixtral 8x7B vs. GPT-4



Improving Data Coverage

Utilizing Reference Solutions

Benchmarks such as MATH and GSM8K come with reference text solutions

Can we use these reference solutions to aid synthetic solution generation?

Utilizing Reference Solutions

Benchmarks such as MATH and GSM8K come with reference text solutions

Can we use these reference solutions to aid synthetic solution generation?

Prompting the model with {Question, Reference Solution} leads to increased training set coverage!

Utilizing Reference Solutions

Benchmarks such as MATH and GSM8K come with reference text solutions

Can we use these reference solutions to aid synthetic solution generation?

Prompting the model with {Question, Reference Solution} leads to increased training set coverage!

The generated solutions can copy the intermediate computation or answer from the text solution

```
The answer is ANSWER
```

Utilizing Reference Solutions by Masking Them

Question

Lynne bought 7 books about cats and 2 books about the solar system. She also bought 3 magazines. Each book cost $7 and each magazine cost $4. How much did Lynne spend in all?

Ground-Truth Text Solution

Lynne bought a total of 7 + 2 = 9 books. The books cost Lynne 9 x 7 = $63. For 3 magazines, Lynne spent 3 x 4 = $12. In total, Lynne spent 63 + 12 = $75

Masked Text Solution

Lynne bought a total of 7 + 2 = M books. The books cost Lynne M x 7 = N. For 3 magazines, Lynne spent 3 x 4 = P. In total, Lynne spent N + P = Q

OpenMathInstruct-1 vs. Other Datasets

Dataset	Size	Generating LM (Permissive License)
MathInstruct (Yue et al. 2024)	262K	GPT-4 (✗)
MetaMathQA (Yu et al. 2024)	395K	GPT-3.5 (✗)
MathCodeInstruct (Wang et al. 2024)	80K	GPT-4 (✗)
WizardMath (Luo et al. 2023)	96K	GPT-3.5 (✗)
ToRA (Gou et al. 2024)	16K	GPT-4 (✗)
OpenMathInstruct-1	1.8M	Mixtral (✓)

Results

Size	Model	GSM8K	MATH
7B	MAmmoTH	59.4	33.4
	ToRA	72.6	**44.6**
	OpenMath-CodeLlama	75.9	43.6
	MetaMath-Mistral-7B	77.7	28.2
	MAmmoTH-7B-Mistral	75.0	40.0
	OpenMath-Mistral-7B	**80.2**	44.5
70B	MetaMath	82.3	26.6
	MAmmoTH	76.9	41.8
	ToRA	84.3	49.7
	OpenMath-Llama2	**84.7**	46.3
	OpenMath-CodeLlama	84.6	**50.7**

OpenMathInstruct-1 models perform on par with the best GPT-4 distilled models!

Conclusion

We introduce OpenMathInstruct-1 in this paper:

- With 1.8 million QA pairs, it is at least four times bigger than prior work
- Strong finetuning results, which are on par or better than the *GPT-distilled* models
- The dataset is released with a commercially permissive license