



OpenMathInstruct-1: A 1.8 Million Math Instruction Tuning Dataset



Shubham Toshniwal



Ivan Moshkov



Sean Narenthiran



Daria Gitman



Fei Jia



Igor Gitman



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 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 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 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:

$$\mathcal{I} (q_1, c_1), \dots, (q_k, c_k) \quad q'$$

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:

$$\mathcal{I} (q_1, c_1), \dots, (q_k, c_k) q'$$

\mathcal{I} represents the task instruction

$\{q_1, \dots, q_k\}$ represent k problems representative of the dataset

$\{c_1, \dots, c_k\}$ represent their respective solutions

q' represents a question from the training set

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:

$$\mathcal{I} (q_1, c_1), \dots, (q_k, c_k) q'$$

\mathcal{I} represents the task instruction

$\{q_1, \dots, q_k\}$ represent k problems representative of the dataset

$\{c_1, \dots, c_k\}$ represent their respective solutions

q' represents a question from the training set

Filter solutions s' that lead to ground truth answer: $\text{ANSWER}(s') = a'$

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:

$$\mathcal{I} (q_1, c_1), \dots, (q_k, c_k) \quad q'$$

\mathcal{I} represents the task instruction

$\{q_1, \dots, q_k\}$ represent k problems representative of the dataset

$\{c_1, \dots, c_k\}$ represent their respective solutions

q' represents a question from the training set

Filter solutions s' that lead to ground truth answer: $\text{ANSWER}(s') = a'$

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

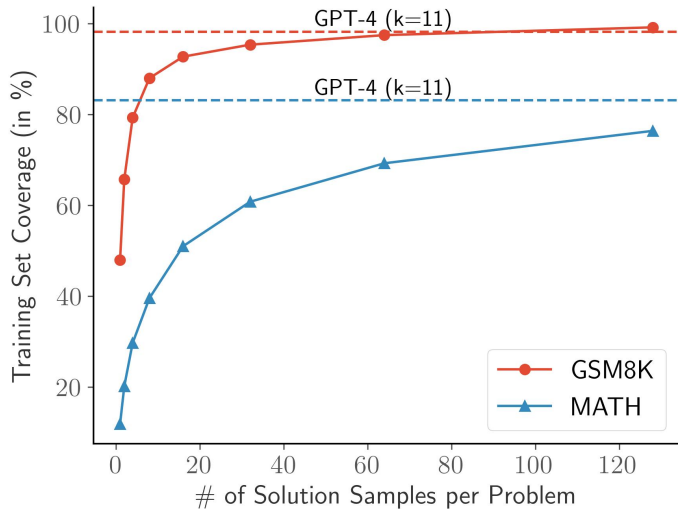
Abstract

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!

$$\mathcal{I} (q_1, t_1, c_1), \dots, (q_k, t_k, c_k) \quad q', t'$$

where t_i 's and t' represent the reference text solutions

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!

$$\mathcal{I} (q_1, t_1, c_1), \dots, (q_k, t_k, c_k) \quad q', t'$$

where t_i 's and t' represent the reference text solutions

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

```
The answer is ANSWER
```

```
return 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?

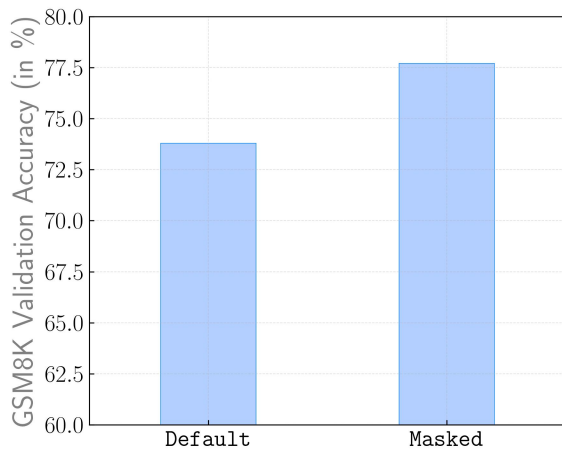
Ground-Truth Text Solution

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

Masked Text Solution

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

Default vs Masked Prompting



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 (✓)

OpenMathInstruct-1: Diversity of Solutions

Question

Let $t(x) = \sqrt{3x+1}$ and $f(x) = 5 - t(x)$. What is $t(f(5))$?

Solution 1

To find $t(f(5))$ we can substitute $x = 5$ into the functions $t(x)$ and $f(x)$ and then find the value of $t(f(5))$. To do the calculations we'll use sympy library.

```
<llm-code>
from sympy import symbols,
sqrt, simplify
```

```
# Define the symbols
x, f = symbols('x f')
```

```
# Define the functions
t = sqrt(3*x + 1)
f = 5 - t
```

```
# Evaluate t at f(5)
t.subs(x, f.subs(x, 5))
</llm-code>
```

```
<llm-code-output>
2
</llm-code-output>
```

So the value of $t(f(5))$ is $\boxed{2}$.

Solution 2

First let's calculate $t(x)$ for $x = 5$:

$$t(5) = \sqrt{3 * 5 + 1} = \sqrt{16} = 4.$$

Then let's calculate $f(x)$ for $x = 5$:

$$f(5) = 5 - t(5) = 5 - 4 = 1.$$

Finally let's calculate

$$t(f(5)) = t(1) = \sqrt{3 * 1 + 1} = \sqrt{4} = 2.$$

So the answer is $\boxed{2}$.

Solution 3

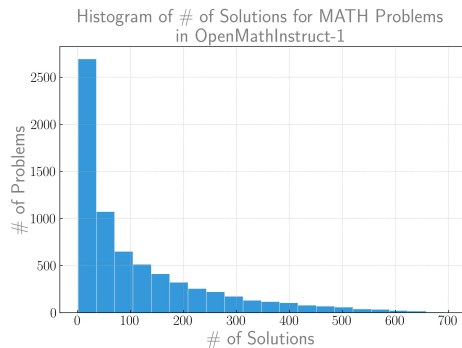
```
<llm-code>
def t(x):
    return (3 * x + 1)
    ** 0.5
```

```
def f(x):
    return 5 - t(x)
```

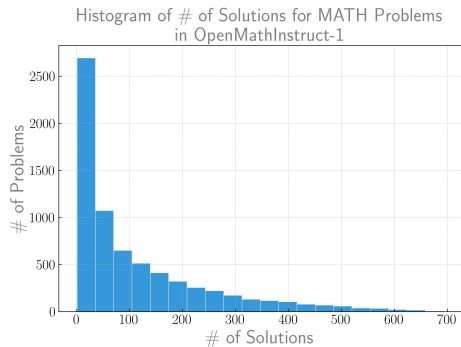
```
t(f(5))
</llm-code>
<llm-code-output>
2.0
</llm-code-output>
```

Thus the answer is $\boxed{2}$.

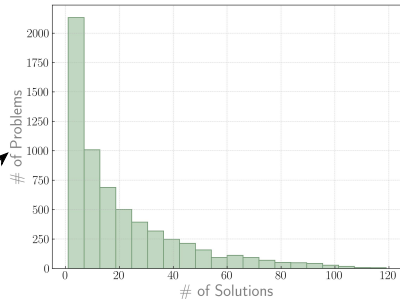
Data Selection



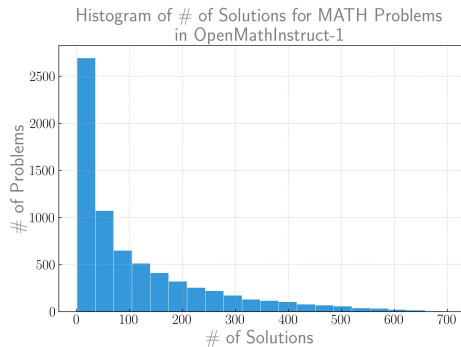
Data Selection



Uniform Sampling

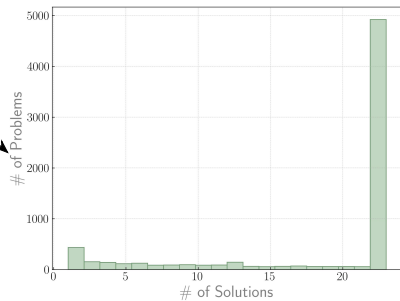
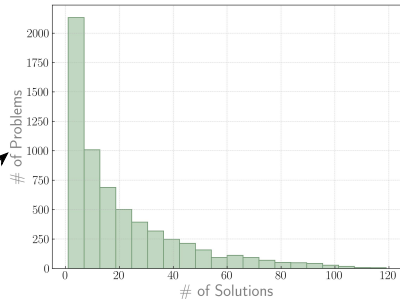


Data Selection

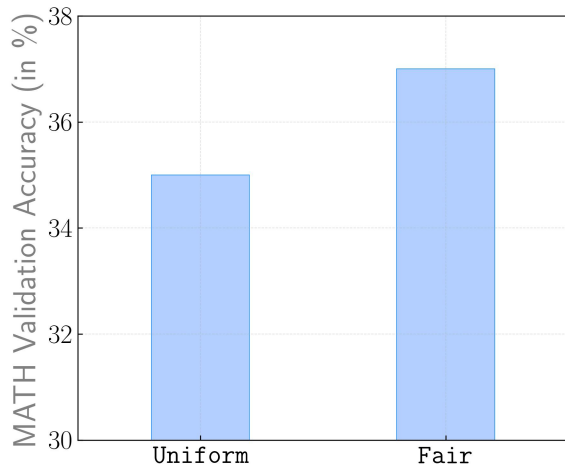


Uniform Sampling

Fair Sampling



Uniform vs. Fair Sampling



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	<u>44.5</u>
70B	MetaMath	82.3	26.6
	MAmmoTH	76.9	41.8
	ToRA	84.3	49.7
	OpenMath-Llama2	84.7	46.3
	OpenMath-CodeLlama	<u>84.6</u>	50.7

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	<u>44.5</u>
70B	MetaMath	82.3	26.6
	MAmmoTH	76.9	41.8
	ToRA	84.3	49.7
	OpenMath-Llama2	84.7	46.3
	OpenMath-CodeLlama	<u>84.6</u>	50.7

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



Home About Participate Team Updates Contact

New: [Second Progress Prize Launched on Kaggle](#)

[← Updates](#)

Resources for Participants

Datasets

United States Competitions (Lower to Higher Difficulty)

American Mathematics Competition 10 (AMC '10) past competition questions: [Website](#)

American Mathematics Competition 12 (AMC '12) past competition questions: [Website](#)

American Invitational Maths Exam (AIME): [Website](#)

UK Competitions (Lower to Higher Difficulty)

UK First Mathematics Challenge: [Website](#)

UK Primary Mathematics Challenge: [Website](#)

UKMT Challenges: [Senior](#), [Intermediate](#) and [Others](#)

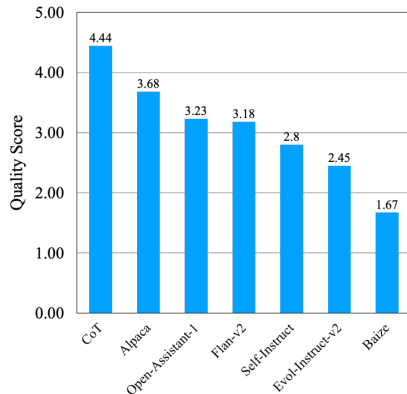
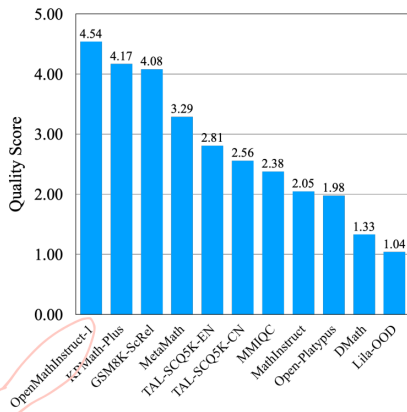
Published Datasets

Hendrycks et al. MATH dataset: [Github](#) | [Paper](#)

OpenMathInstruct-1: [Paper](#)

Suggested by the official AI Mathematical Olympiad (AIMO) website and used by some of the participants in AIMO-Progress Prize 1

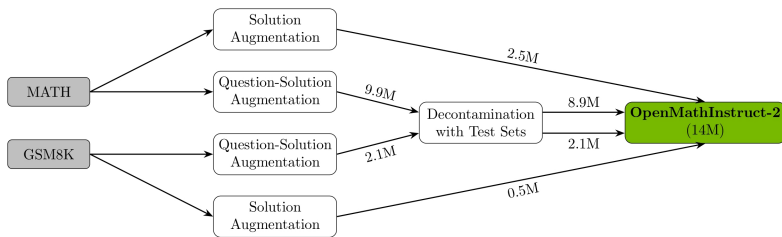
Impact: Quality



(a) Quality scores by Scorer-DeepSeek-Reasoning. (b) Quality scores by Scorer-DeepSeek-General.

Exploring the Mystery of Influential Data for Mathematical Reasoning by Ni et al. identifies [OpenMathInstruct-1](#) as the highest quality math reasoning dataset

Follow-up: OpenMathInstruct-2



Released in October with a commercially permissive license

14M QA pairs: **8x bigger than OpenMathInstruct-1**; includes 592K synthesized questions

[OpenMath2-Llama3.1-8B](#) outperforms [Llama3.1-8B-Instruct](#) on MATH by 15.9%

Used by OLMo2 released by Ai2 in November-end

Solutions with Incorrect Reasoning

Question

The areas of two squares are in the ratio 25 : 36. What is the ratio of their perimeters? Express your answer in the form $a : b$.

Generated Solution

Let's use sympy to print out the difference between the ratio of their perimeters and each of the options.

```
<llm-code>
from sympy import Rational, Abs

# areas are in the ratio 25:36
area_ratio = Rational(25, 36)

# list of options
options = [Rational(5, 6), Rational(5, 4), Rational(5, 3),
          Rational(5, 2), Rational(5, 1)]

# let's print out the differences
[Abs(area_ratio - frac_option) for frac_option in options]
</llm-code>
<llm-code-output>
[5/36, 5/9, 35/36, 65/36, 155/36]
</llm-code-output>
```

Let's now check which difference is the smallest.

```
<llm-code>
import numpy as np

# Calculate the idx of the closest option
min_idx = np.argmin([5/36, 5/9, 35/36, 65/36, 155/36])

# Print the closest option
print(options[min_idx])
</llm-code>
<llm-code-output>
5/6
</llm-code-output>
```

So the answer is .

Solutions with Incorrect Reasoning

Question

The areas of two squares are in the ratio 25 : 36. What is the ratio of their perimeters? Express your answer in the form $a : b$.

Generated Solution

Let's use sympy to print out the difference between the ratio of their perimeters and each of the options.

```
<llm-code>
from sympy import Rational, Abs

# areas are in the ratio 25:36
area_ratio = Rational(25, 36)

# list of options
options = [Rational(5, 6), Rational(5, 4), Rational(5, 3),
          Rational(5, 2), Rational(5, 1)]

# let's print out the differences
[Abs(area_ratio - frac_option) for frac_option in options]
</llm-code>
<llm-code-output>
[5/36, 5/9, 35/36, 65/36, 155/36]
</llm-code-output>
```

Let's now check which difference is the smallest.

```
<llm-code>
import numpy as np

# Calculate the idx of the closest option
min_idx = np.argmin([5/36, 5/9, 35/36, 65/36, 155/36])

# Print the closest option
print(options[min_idx])
</llm-code>
<llm-code-output>
5/6
</llm-code-output>
```

So the answer is .

Hallucinated list
of options

Solutions with Incorrect Reasoning

Question

The areas of two squares are in the ratio 25 : 36. What is the ratio of their perimeters? Express your answer in the form $a : b$.

Generated Solution

Let's use sympy to print out the difference between the ratio of their perimeters and each of the options.

```
<llm-code>
from sympy import Rational, Abs

# areas are in the ratio 25:36
area_ratio = Rational(25, 36)

# list of options
options = [Rational(5, 6), Rational(5, 4), Rational(5, 3),
          Rational(5, 2), Rational(5, 1)]
```

```
# let's print out the differences
[Abs(area_ratio - frac_option) for frac_option in options]
</llm-code>
<llm-code-output>
[5/36, 5/9, 35/36, 65/36, 155/36]
</llm-code-output>
```

Let's now check which difference is the smallest.

```
<llm-code>
import numpy as np

# Calculate the idx of the closest option
min_idx = np.argmin([5/36, 5/9, 35/36, 65/36, 155/36])

# Print the closest option
print(options[min_idx])
</llm-code>
<llm-code-output>
5/6
</llm-code-output>
```

So the answer is 5 : 6.

Hallucinated list
of options

Irrelevant Reasoning

Solutions with Incorrect Reasoning

Question

The areas of two squares are in the ratio 25 : 36. What is the ratio of their perimeters? Express your answer in the form $a : b$.

Generated Solution

Let's use sympy to print out the difference between the ratio of their perimeters and each of the options.

```
<llm-code>
from sympy import Rational, Abs

# areas are in the ratio 25:36
area_ratio = Rational(25, 36)

# list of options
options = [Rational(5, 6), Rational(5, 4), Rational(5, 3),
          Rational(5, 2), Rational(5, 1)]
```

```
# let's print out the differences
[Abs(area_ratio - frac_option) for frac_option in options]
</llm-code>
<llm-code-output>
[5/36, 5/9, 35/36, 65/36, 155/36]
</llm-code-output>
```

Let's now check which difference is the smallest.

```
<llm-code>
import numpy as np

# Calculate the idx of the closest option
min_idx = np.argmin([5/36, 5/9, 35/36, 65/36, 155/36])

# Print the closest option
print(options[min_idx])
</llm-code>
<llm-code-output>
5/6
</llm-code-output>
```

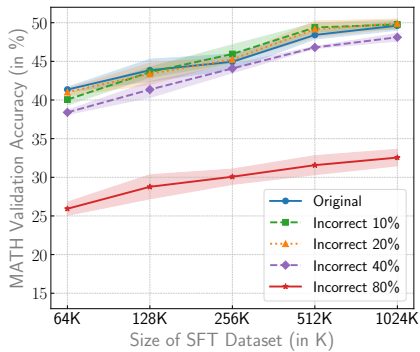
So the answer is 5 : 6.

Hallucinated list
of options

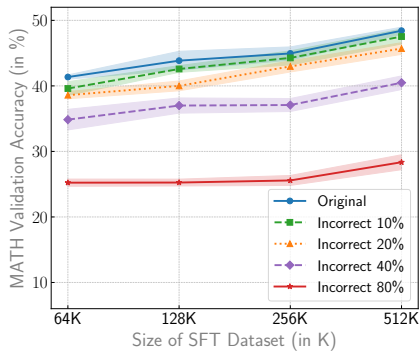
Irrelevant Reasoning

Correct Answer!

OpenMathInstruct-2: Robustness of SFT



Correct solutions mismatched with questions



Adding wrong-answer solutions

Conducted extensive experiments on removing and adding noisy data to the SFT blend

Conclusion: SFT performance is robust to presence of up-to 20% low-quality data

Conclusion

We introduce OpenMathInstruct-1 in this paper:

With 1.8M 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

The dataset has been widely used by the community



Conclusion

We introduce OpenMathInstruct-1 in this paper:

With 1.8M 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

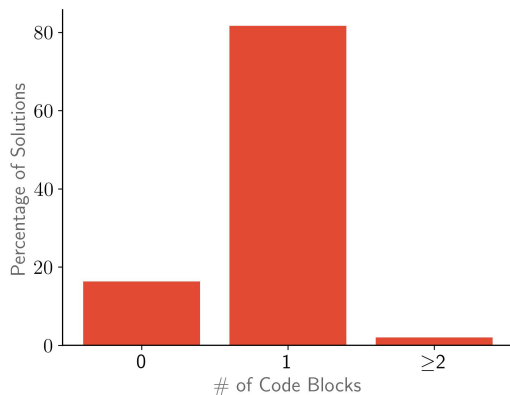
The dataset has been widely used by the community

Recently released a follow-up, [OpenMathInstruct-2](#), which is 8x bigger than OpenMathInstruct-1



Backup Data Slides

Code-Block Count Frequencies



Distribution of solutions with different code-block counts in OpenMathInstruct-1.

Error Analysis

Error Type	Count
Text Reasoning Error	189
Code Reasoning Error	292
Code Execution Error	78
Code timeout	15
Max code executions reached	10
Total	584

Impact of Code-Preferential Data Selection

Prompt	Pass@1	SC (k=4)
Default	37.4	45.2
Majority-Code	39.8	42.6
Any-Code	39.4	42.6