ACL 2026 SAN DIEGO JULY 2-7

OlymMATH

Challenging the Boundaries of Reasoning: An Olympiad-Level Math Benchmark for LLMs

Haoxiang Sun `hxiang-sun.github.io`

Renmin University of China

with Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, and Ji-Rong Wen

AI Box

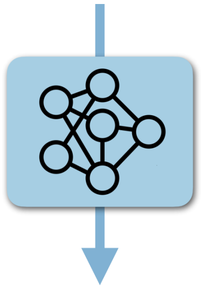
Background: Math Reasoning with LLMs

Background: Math Reasoning with LLMs

Input: Sixteen is 64% of what number?

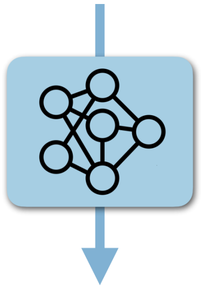
Background: Math Reasoning with LLMs

Input: Sixteen is 64% of what number?



Background: Math Reasoning with LLMs

Input: Sixteen is 64% of what number?

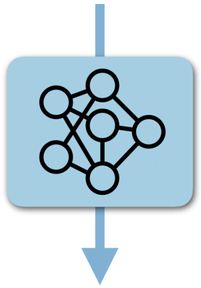


Output:

<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Background: Math Reasoning with LLMs

Input: Sixteen is 64% of what number?

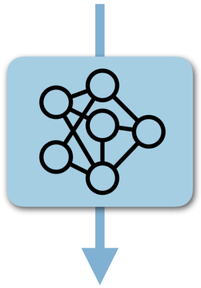


Output:

<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{**25**}.

Background: Math Reasoning with LLMs

Input: Sixteen is 64% of what number?

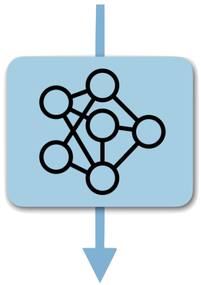


Output:

<think>If it is x, we have $0.64x = 16$, thus $x = 25$.</think> The answer is \boxed{**25**}.

Background: Math Reasoning with LLMs

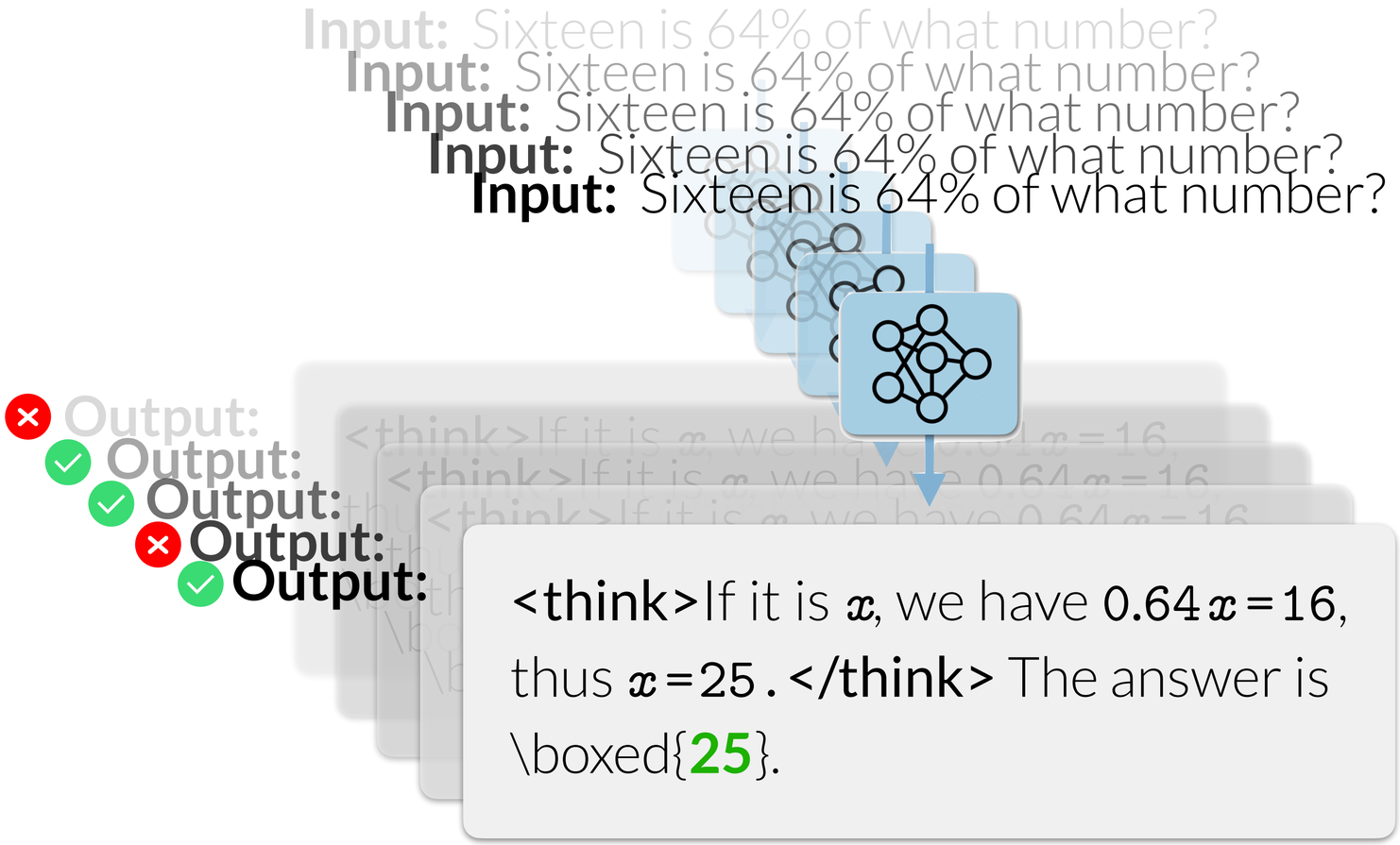
Input: Sixteen is 64% of what number?



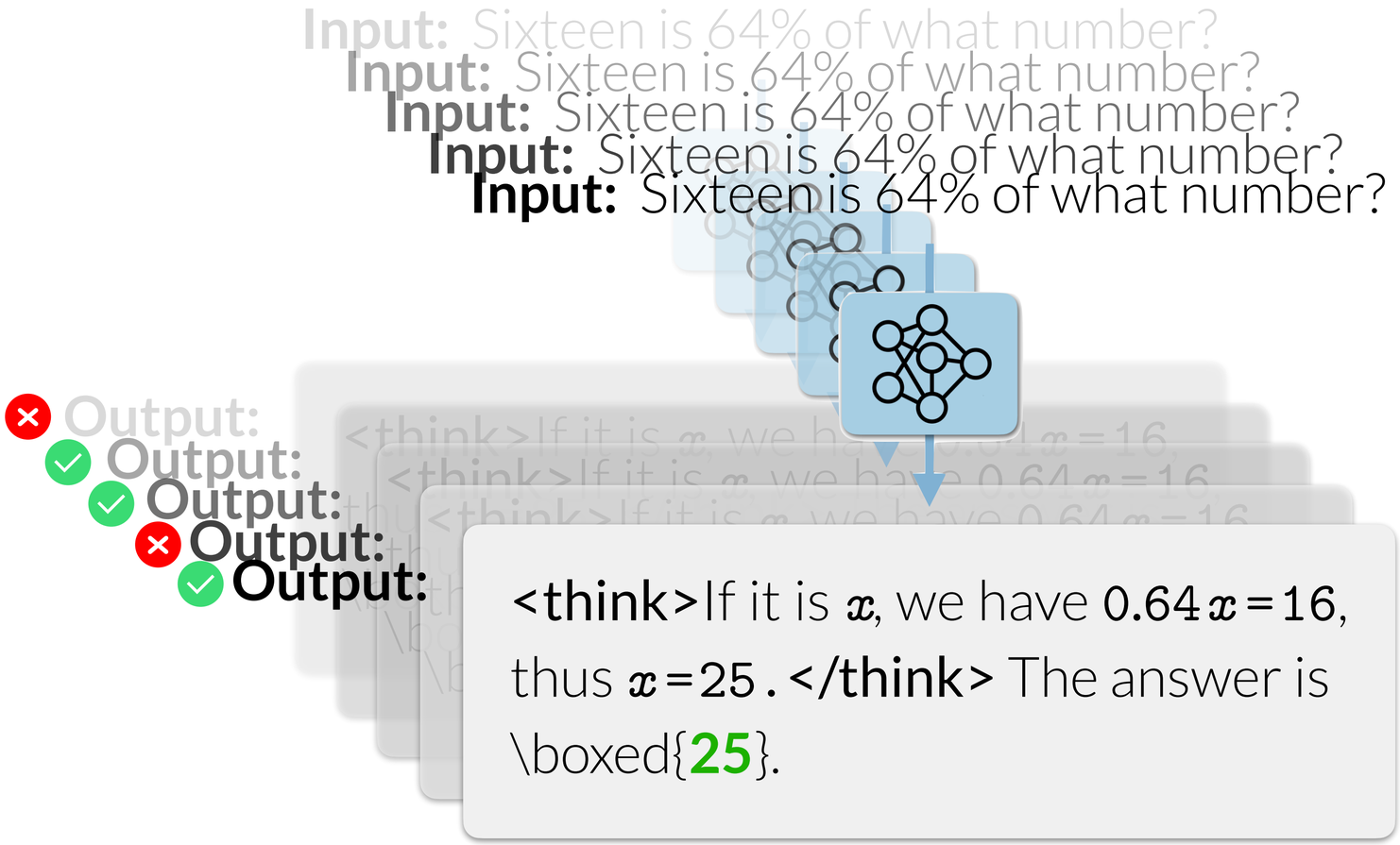
Output:

<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{**25**}.

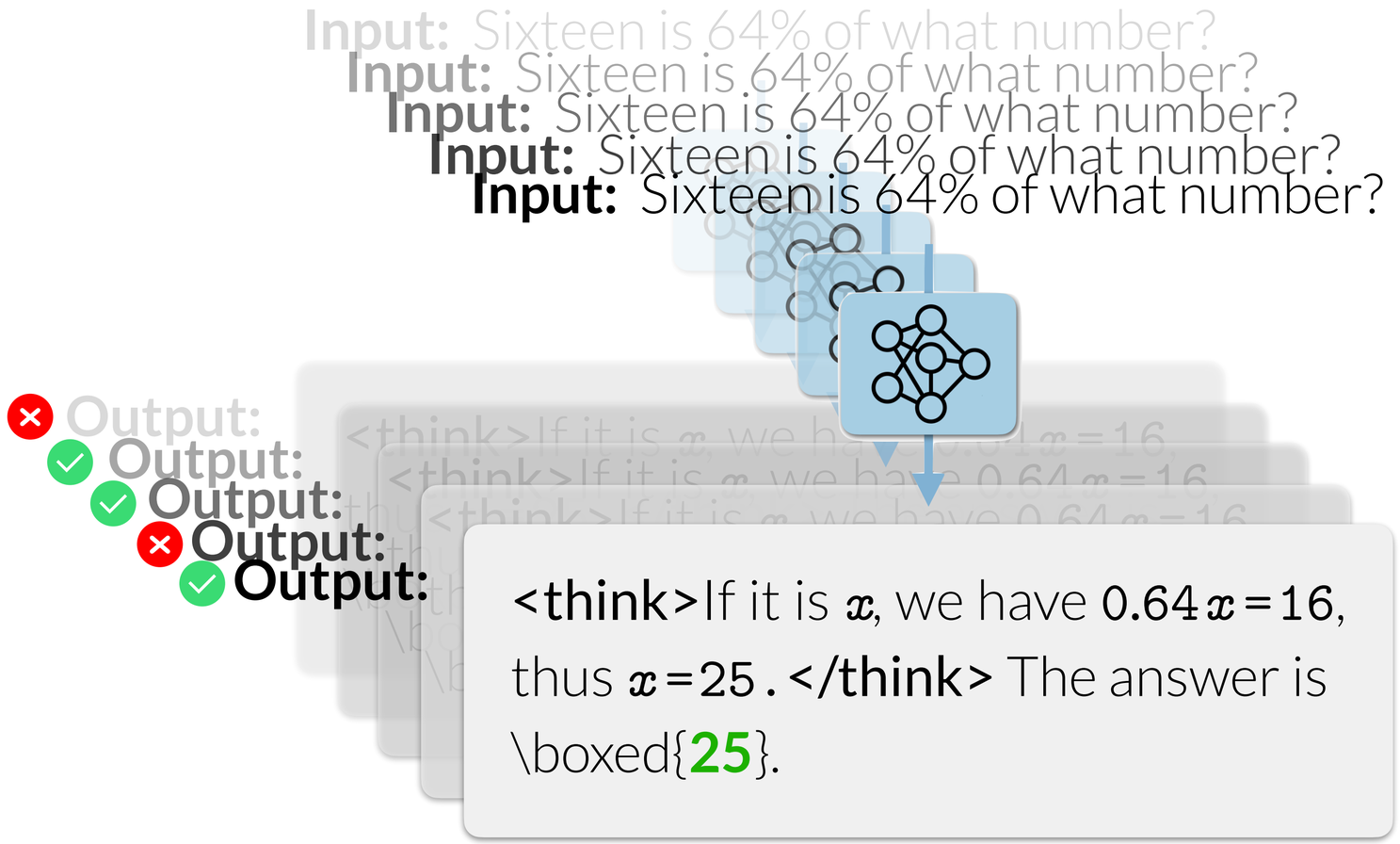
Background: Math Reasoning with LLMs



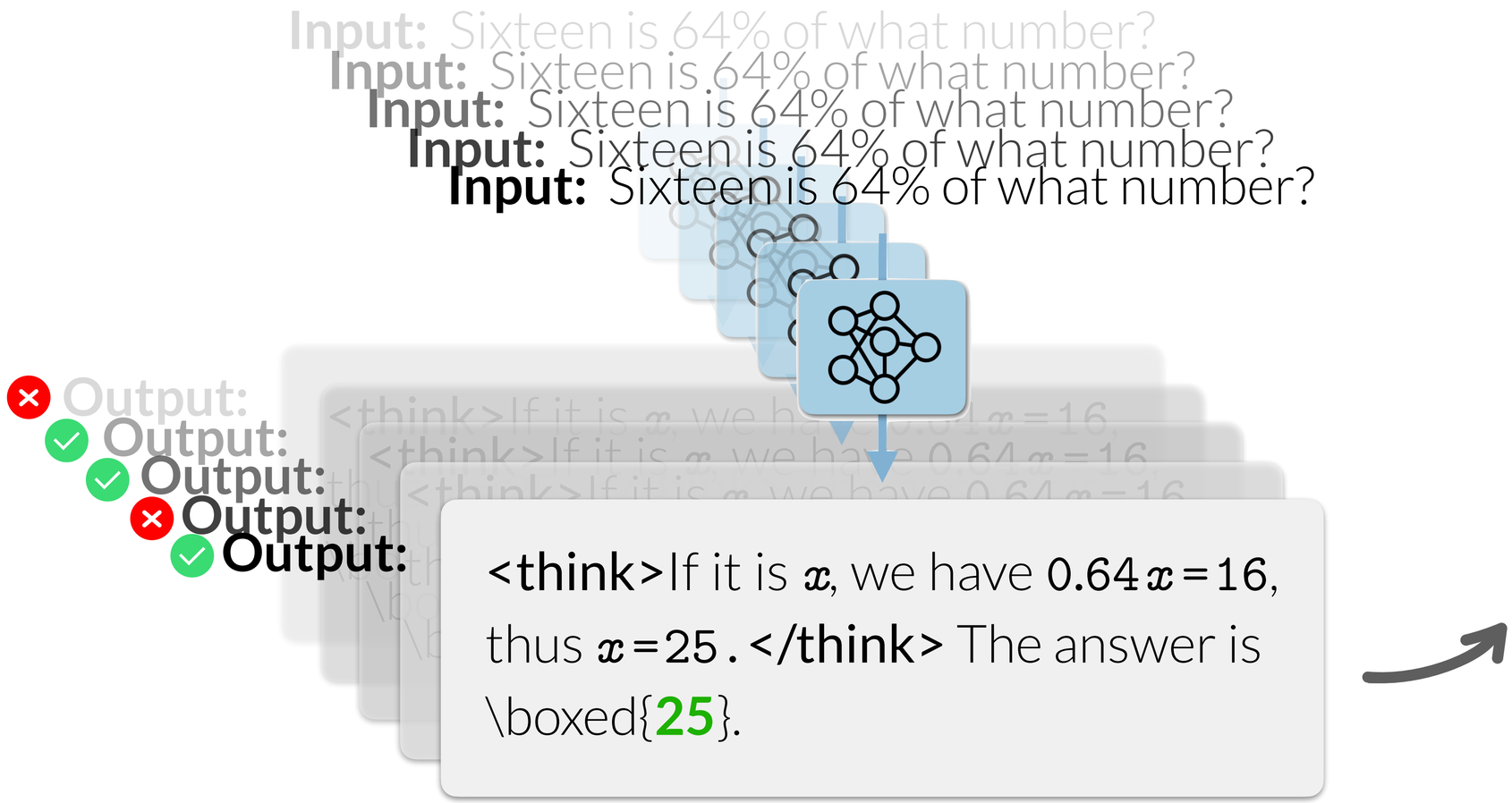
Background: Math Reasoning with LLMs



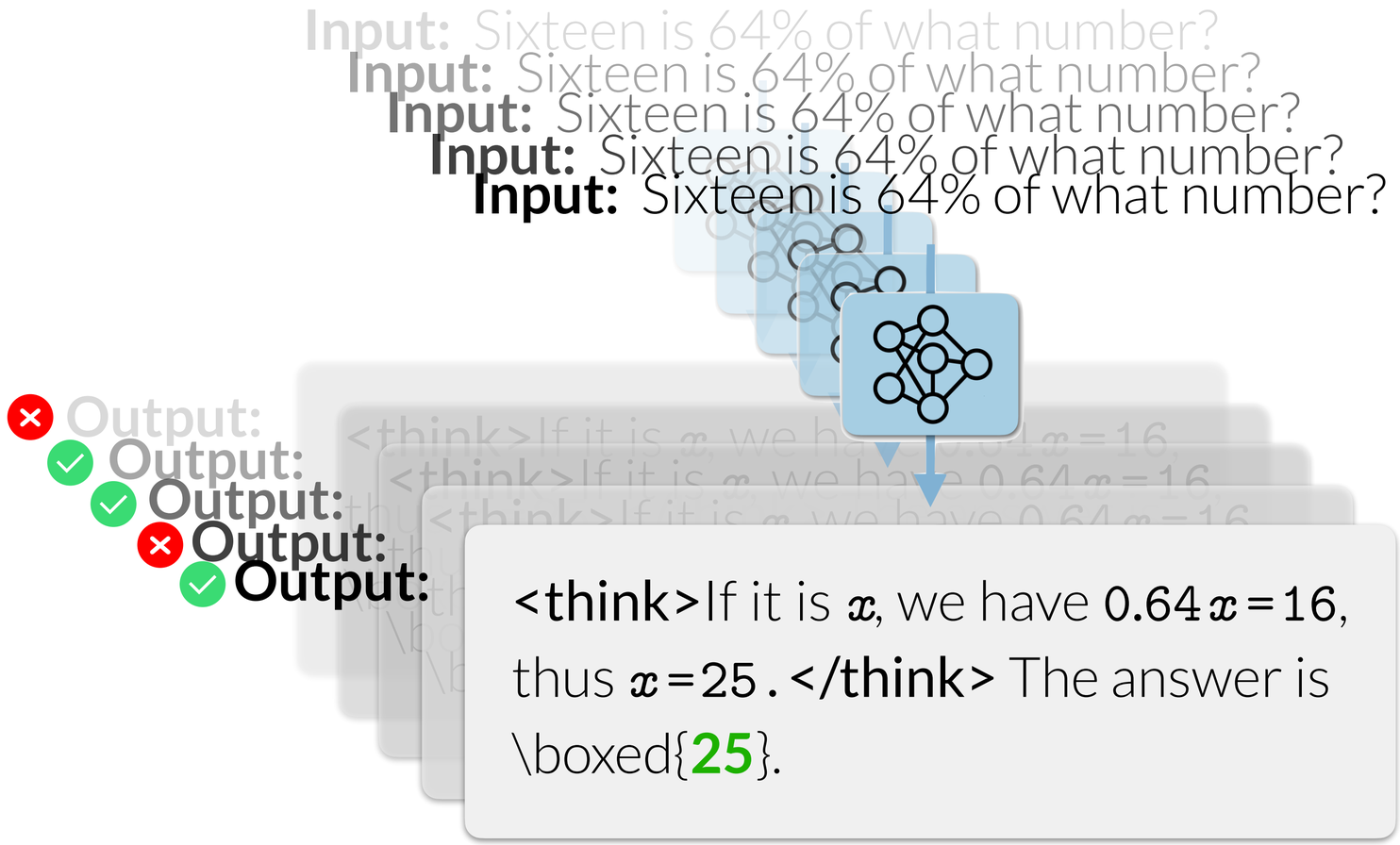
Evaluating LLM's Math Reasoning



Evaluating LLM's Math Reasoning



Evaluating LLM's Math Reasoning



Pass@1: 0.6

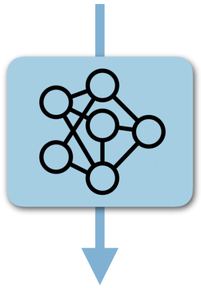
Pass@5: 1.0

Cons@5: 1.0

Motivation

Motivation

Input: Sixteen is 64% of what number?

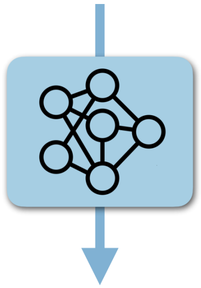


Output:

<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Motivation

Input: **Sixteen is 64% of what number?**

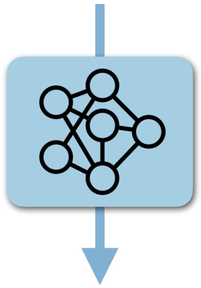


Output:

<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Motivation

Input: **Sixteen is 64% of what number?**



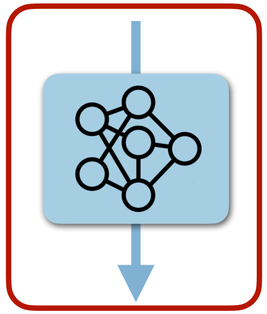
Output:

<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Challenging Enough?

Motivation

Input: Sixteen is 64% of what number?



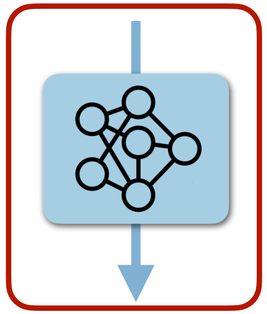
Output:

<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Challenging Enough?

Motivation

Input: Sixteen is 64% of what number?



Output:

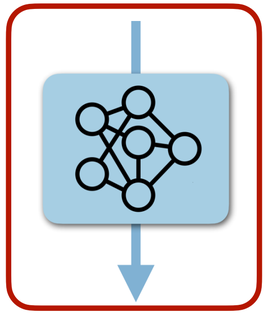
<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Challenging Enough?

Reliable Enough?

Motivation

Input: Sixteen is 64% of what number?



- Data Leakage

Output:

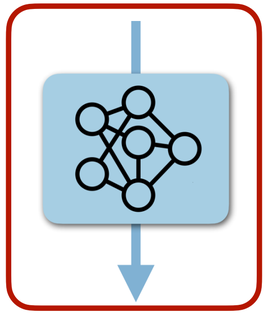
<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Challenging Enough?

Reliable Enough?

Motivation

Input: Sixteen is 64% of what number?



- Data Leakage
- Multilingual

Output:

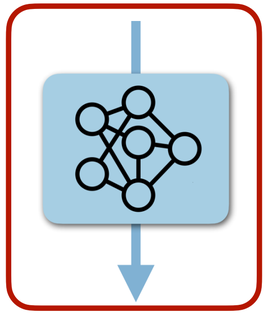
<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Challenging Enough?

Reliable Enough?

Motivation

Input: Sixteen is 64% of what number?



- Data Leakage
- Multilingual
- Statistical Stability

Output:

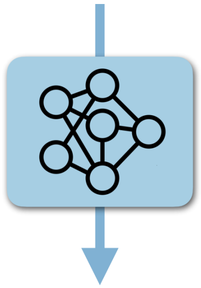
<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Challenging Enough?

Reliable Enough?

Motivation

Input: Sixteen is 64% of what number?



- Data Leakage
- Multilingual
- Statistical Stability

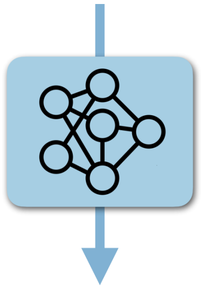
Output:

<think>**If it is x, we have $0.64x=16$, thus $x=25$.** </think> The answer is \boxed{**25**}.

Challenging Enough? Reliable Enough?

Motivation

Input: Sixteen is 64% of what number?



- Data Leakage
- Multilingual
- Statistical Stability

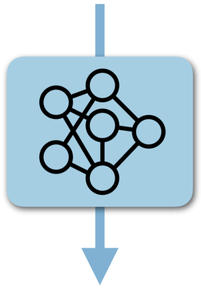
Output:

<think>**If it is x, we have $0.64x=16$, thus $x=25$.** </think> The answer is \boxed{**25**}.

Challenging Enough? Reliable Enough? **Rigorous Enough?**

Motivation

Input: Sixteen is 64% of what number?



- Data Leakage
- Multilingual
- Statistical Stability

Output:

<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Challenging Enough? **Reliable Enough?** **Rigorous Enough?**

Overview:

Overview: OlymMATH

Overview: OlymMATH

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016\times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

Overview: OlymMATH

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

OlymMATH-EASY

Overview: OlymMATH

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016\times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

OlymMATH-EASY

OlymMATH-HARD

Overview: OlymMATH

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

OlymMATH-EASY	OlymMATH-HARD
100 Problems	100 Problems
AIME Level	More Challenging

Overview: OlymMATH

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

Subject: Number Theory. **Formal Statement:**

```
theorem to_prove
  (n : Nat) (p : Nat) (hp : p.Prime)
  (hdiv : p | 2^n + 1) : p % 8 != 7 := by sorry
```

OlymMATH-EASY

100 Problems

AIME Level

OlymMATH-HARD

100 Problems

More Challenging

Overview: OlymMATH

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

Subject: Number Theory. **Formal Statement:**

```
theorem to_prove
  (n : Nat) (p : Nat) (hp : p.Prime)
  (hdiv : p | 2^n + 1) : p % 8 != 7 := by sorry
```

OlymMATH-EASY

100 Problems

AIME Level

OlymMATH-HARD

100 Problems

More Challenging

OlymMATH-LEAN

Overview: OlymMATH

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

Subject: Number Theory. **Formal Statement:**

```
theorem to_prove
  (n : Nat) (p : Nat) (hp : p.Prime)
  (hdiv : p | 2^n + 1) : p % 8 != 7 := by sorry
```

Lean encodes proofs as programs, so proof correctness reduces to type-checking.

OlymMATH-EASY

100 Problems

AIME Level

OlymMATH-HARD

100 Problems

More Challenging

OlymMATH-LEAN

Overview: OlymMATH

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

Subject: Number Theory. **Formal Statement:**

```
theorem to_prove
  (n : Nat) (p : Nat) (hp : p.Prime)
  (hdiv : p | 2^n + 1) : p % 8 != 7 := by sorry
```

Lean encodes proofs as programs, so proof correctness reduces to type-checking.

OlymMATH-EASY

100 Problems

AIME Level

OlymMATH-HARD

100 Problems

More Challenging

OlymMATH-LEAN

150 Problems

Overview: OlymMATH

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

Subject: Number Theory. **Formal Statement:**

```
theorem to_prove
  (n : Nat) (p : Nat) (hp : p.Prime)
  (hdiv : p | 2^n + 1) : p % 8 != 7 := by sorry
```

Lean encodes proofs as programs, so proof correctness reduces to type-checking.

OlymMATH-EASY

100 Problems

AIME Level

EN / CH

OlymMATH-HARD

100 Problems

More Challenging

EN / CH

OlymMATH-LEAN

150 Problems

Overview: OlymMATH

Lean encodes proofs as programs, so proof correctness reduces to type-checking.

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

Subject: Number Theory. **Formal Statement:**

```
theorem to_prove
  (n : Nat) (p : Nat) (hp : p.Prime)
  (hdiv : p | 2^n + 1) : p % 8 != 7 := by sorry
```

OlymMATH-EASY

100 Problems

AIME Level

EN / CH

OlymMATH-HARD

100 Problems

More Challenging

EN / CH

All manually crafted & verified

OlymMATH-LEAN

150 Problems

LEAN

Multi-agent synthesized

Overview: OlymMATH

Lean encodes proofs as programs, so proof correctness reduces to type-checking.

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

Subject: Number Theory. **Formal Statement:**

```
theorem to_prove
  (n : Nat) (p : Nat) (hp : p.Prime)
  (hdiv : p | 2^n + 1) : p % 8 != 7 := by sorry
```

OlymMATH-EASY

100 Problems

AIME Level

EN / CH

OlymMATH-HARD

100 Problems

More Challenging

EN / CH

All manually crafted & verified

OlymMATH-LEAN

150 Problems

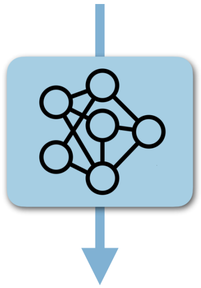
LEAN

Multi-agent synthesized

Please check our paper to see more details on design and evaluation!

Motivation

Input: Sixteen is 64% of what number?



- Data Leakage
- Multilingual
- Statistical Stability

Output:

<think>If it is x, we have $0.64x=16$, thus $x=25$.</think> The answer is \boxed{25}.

Challenging Enough? **Reliable Enough?** **Rigorous Enough?**

Motivation

Challenging Enough?

Reliable Enough?

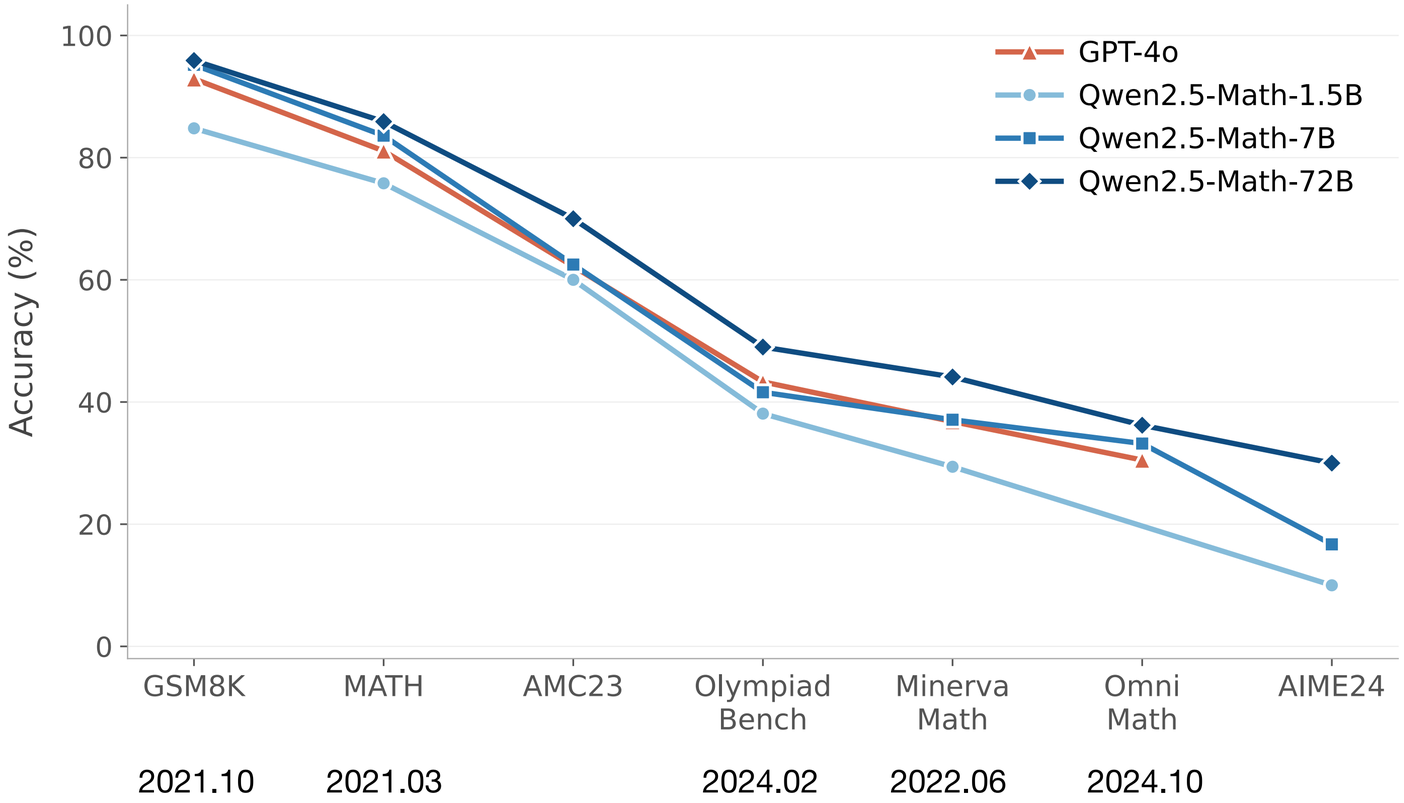
Rigorous Enough?

Motivation

Challenging Enough?

Reliable Enough?

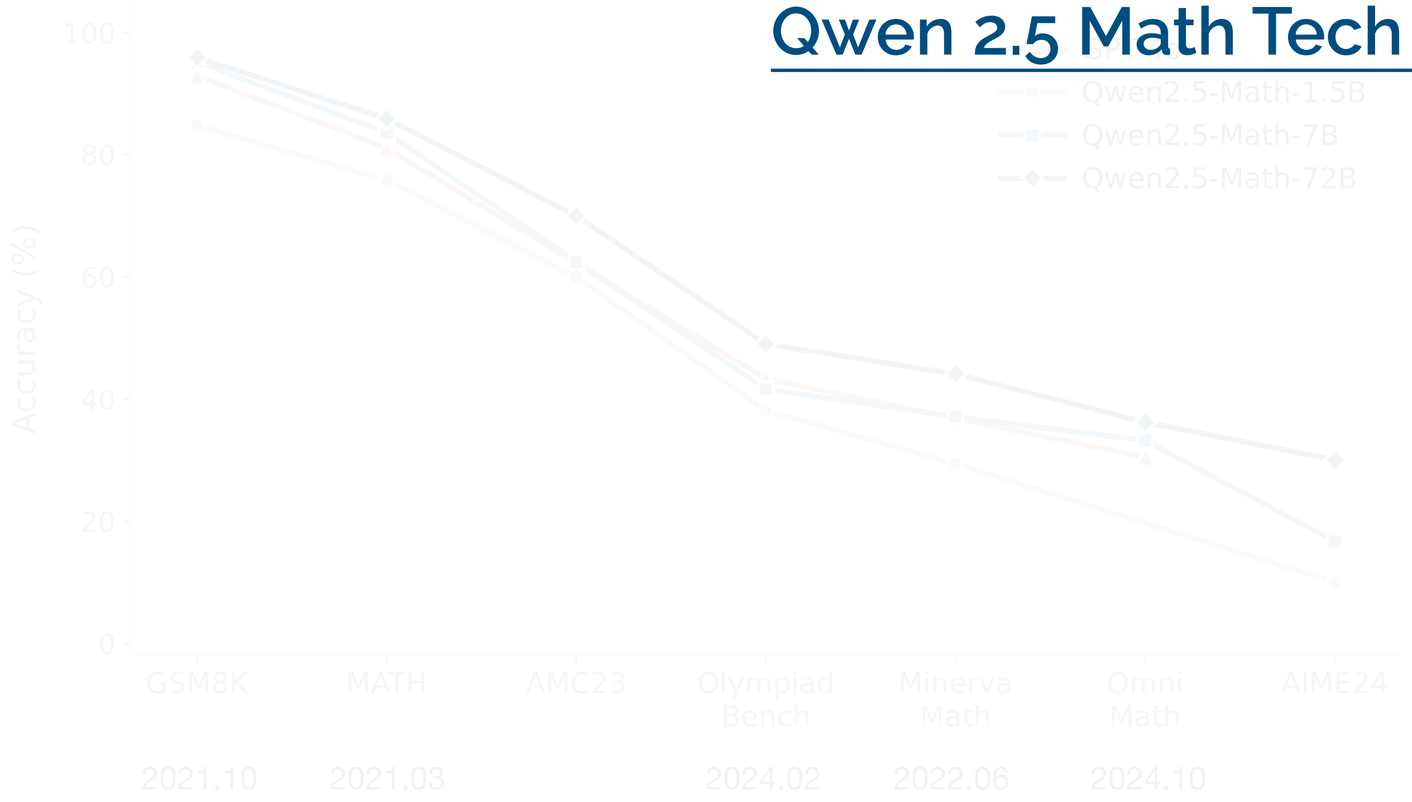
Rigorous Enough?



Motivation

Challenging Enough? Reliable Enough? Rigorous Enough?

Qwen 2.5 Math Tech Report



Motivation

Challenging Enough? Reliable Enough? Rigorous Enough?

Qwen 2.5 Math Tech Report

outperforming NuminaMath-72B CoT. Moreover, Qwen2.5-Math-72B-Instruct nearly achieves a perfect score in TIR mode, solving almost all the problems. We attribute this impressive performance to the extensive amounts of challenging mathematical data collected and synthesized during pre-training. On the extremely difficult AIME 2024 benchmark, Claude3 Opus, GPT-4 Turbo, and Gemini 1.5 Pro manage to solve only 1 or 2 questions out of 30. In contrast, Qwen2.5-Math-72B-Instruct solves 9 problems in Greedy decoding CoT mode and 12 problems in TIR mode. With the help of the RM, Qwen2.5-Math-7B-Instruct could even solve up to 21 problems, further demonstrating the outstanding mathematical problem-solving ability of Qwen2.5-Math-Instruct.

Motivation

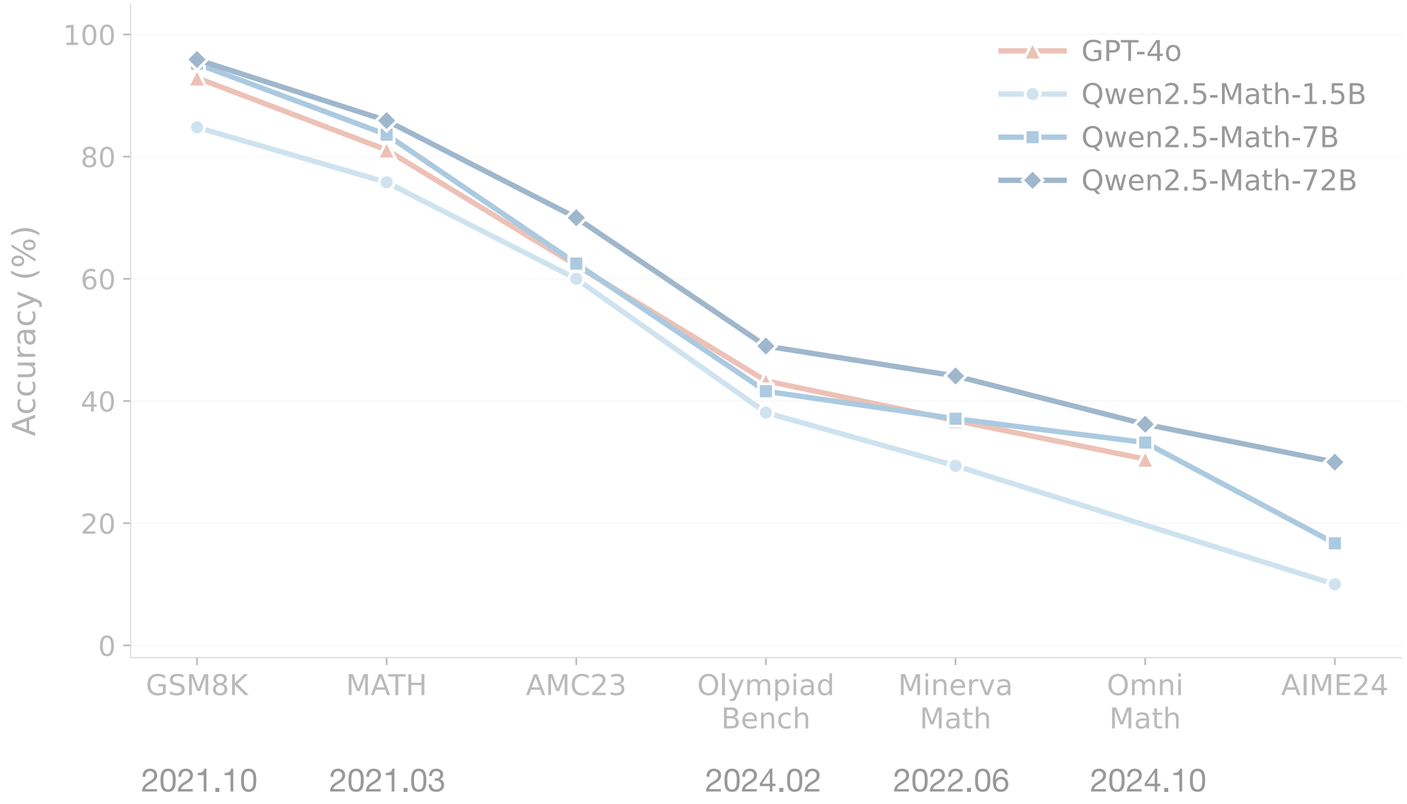
Challenging Enough? Reliable Enough? Rigorous Enough?

Qwen 2.5 Math Tech Report

outperforming NuminaMath-72B CoT. Moreover, Qwen2.5-Math-72B-Instruct nearly achieves a perfect score in TIR mode, solving almost all the problems. We attribute this impressive performance to the extensive amounts of challenging mathematical data collected and synthesized during pre-training. On the extremely difficult AIME 2024 benchmark, Claude3 Opus, GPT-4 Turbo, and Gemini 1.5 Pro manage to solve only 1 or 2 questions out of 30. In contrast, Qwen2.5-Math-72B-Instruct solves 9 problems in Greedy decoding CoT mode and 12 problems in TIR mode. With the help of the RM, Qwen2.5-Math-7B-Instruct could even solve up to 21 problems, further demonstrating the outstanding mathematical problem-solving ability of Qwen2.5-Math-Instruct.

Motivation

Challenging Enough? Reliable Enough? Rigorous Enough?

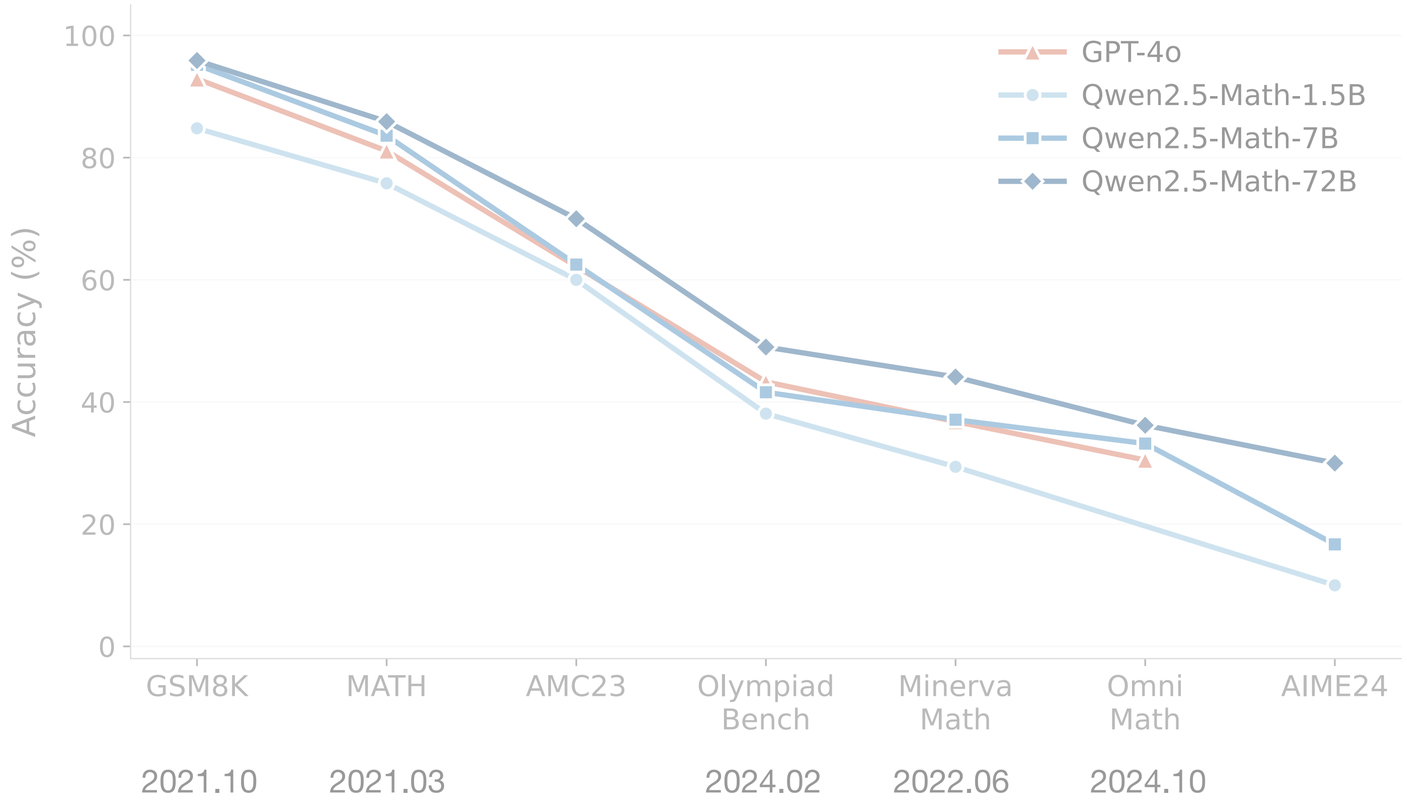


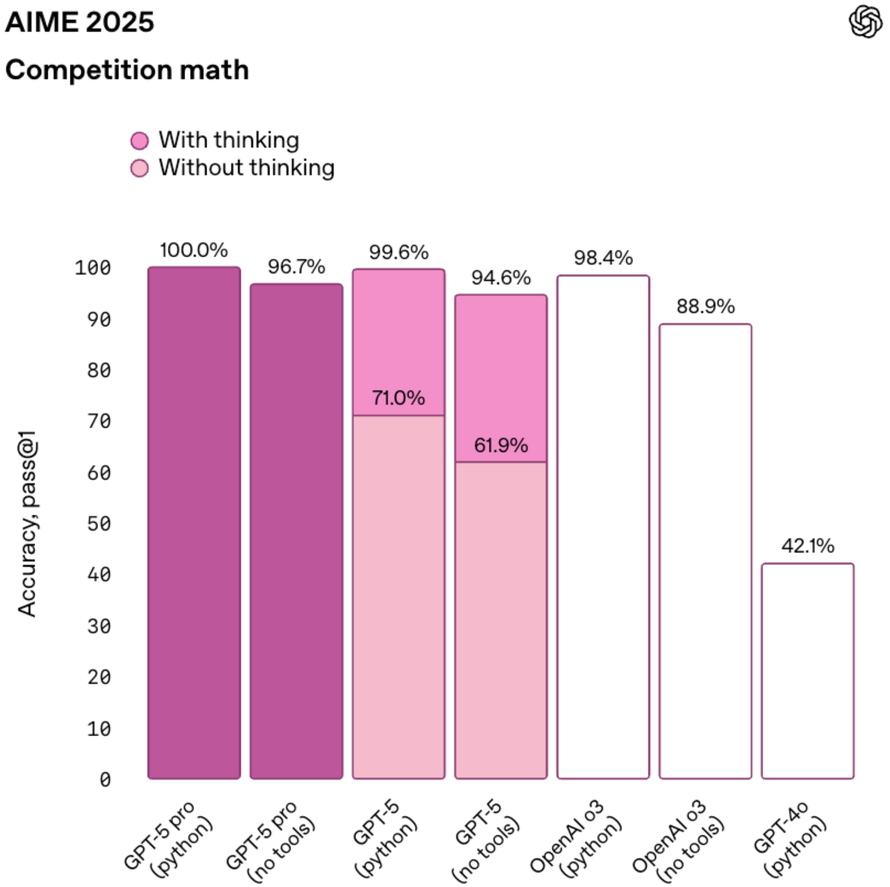
Motivation

Challenging Enough?

Reliable Enough?

Rigorous Enough?



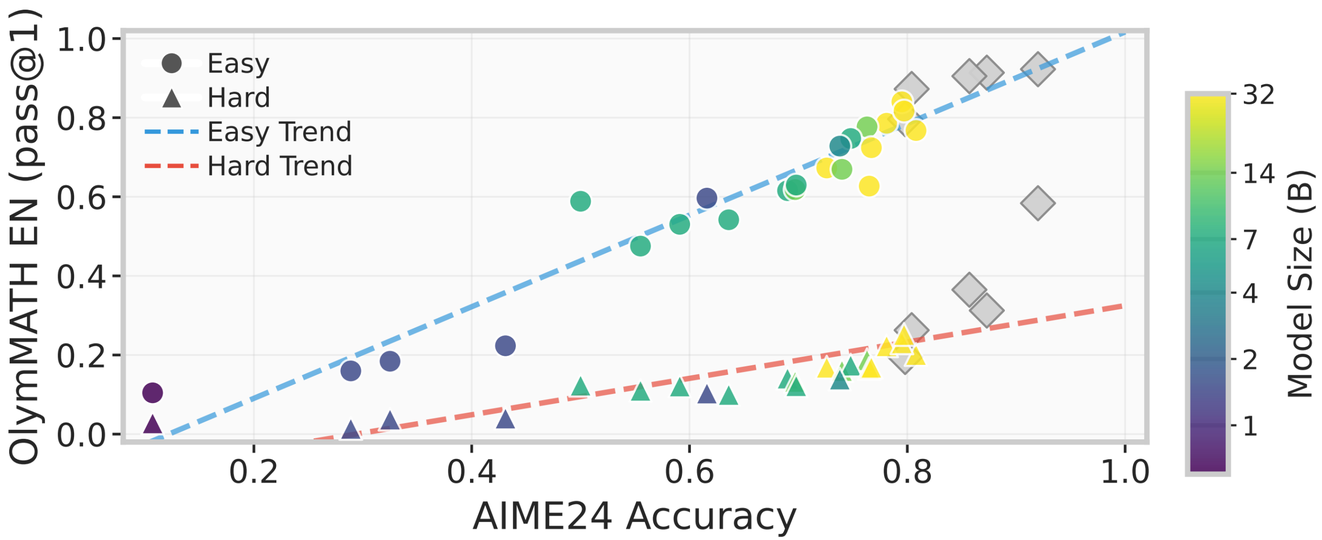


Design

Challenging Enough? Reliable Enough? Rigorous Enough?

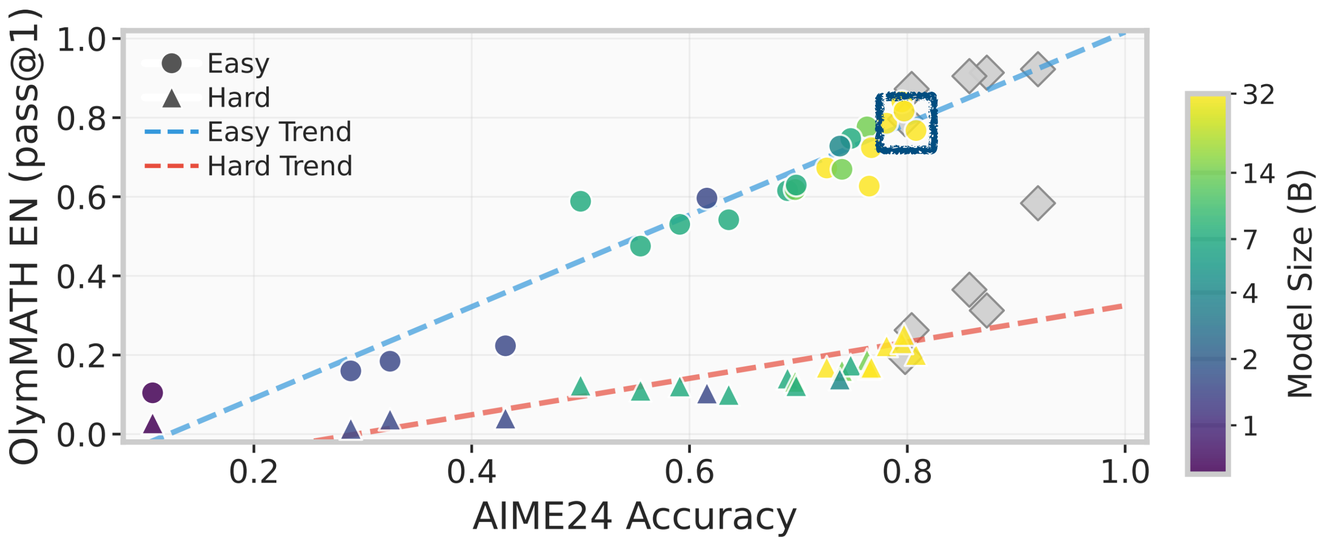
Design

Challenging Enough? Reliable Enough? Rigorous Enough?



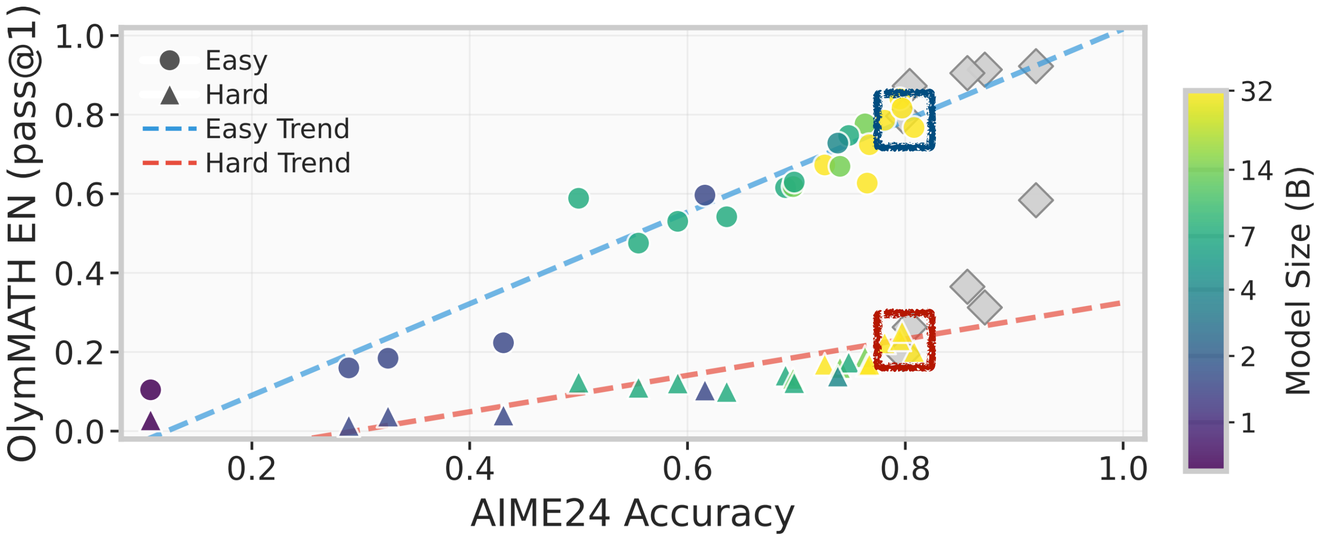
Design

Challenging Enough? Reliable Enough? Rigorous Enough?



Design

Challenging Enough? Reliable Enough? Rigorous Enough?

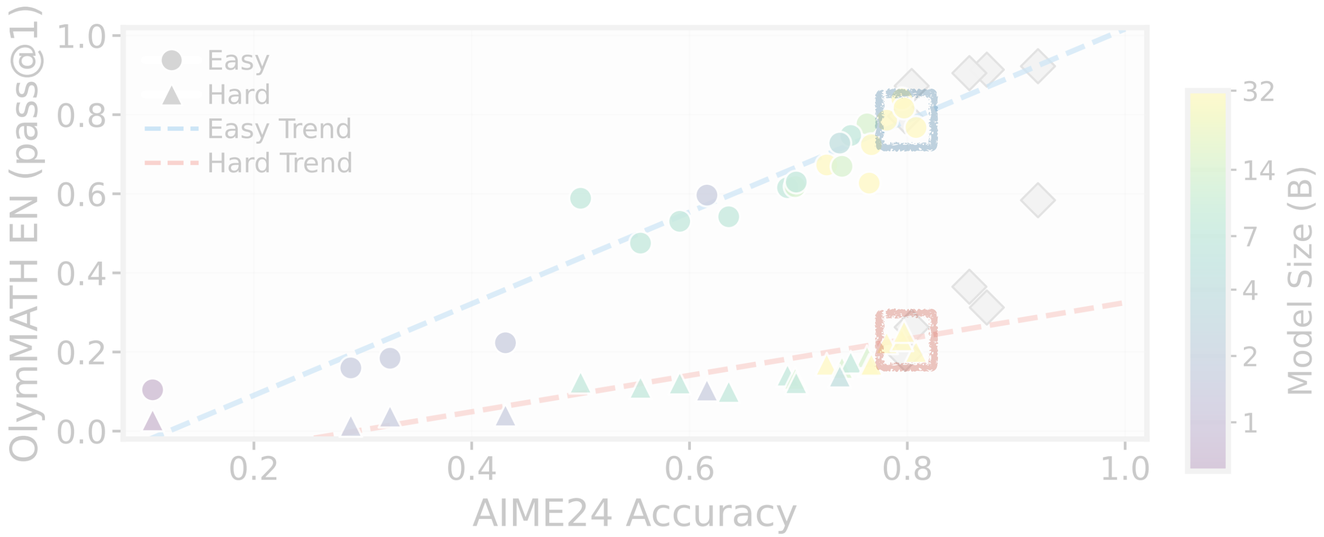


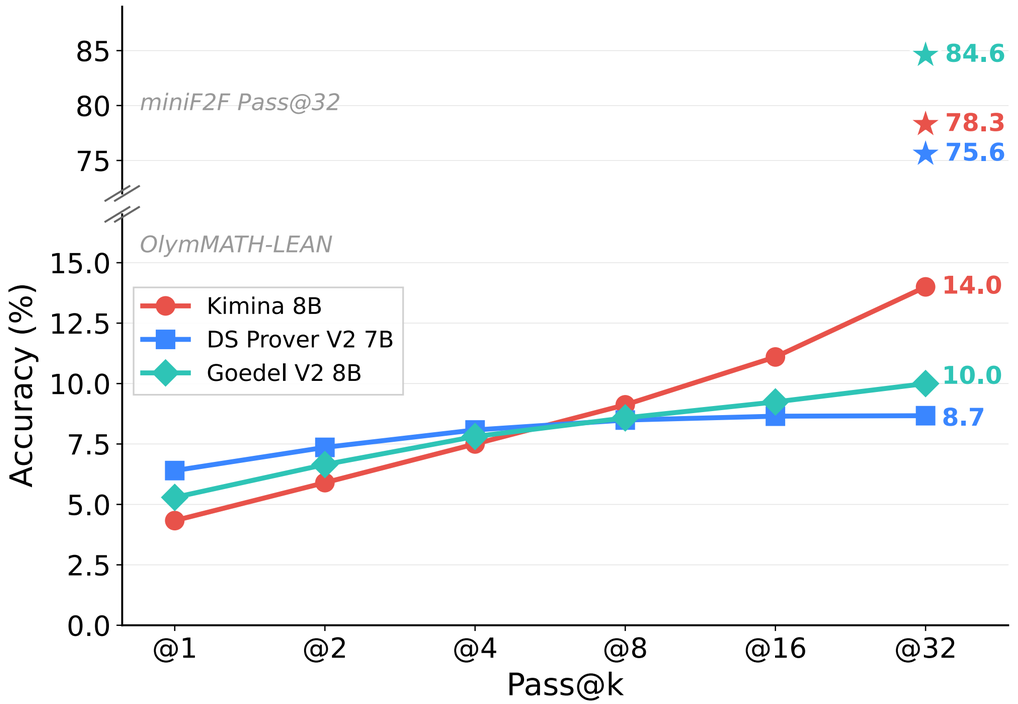
Design

Challenging Enough?

Reliable Enough?

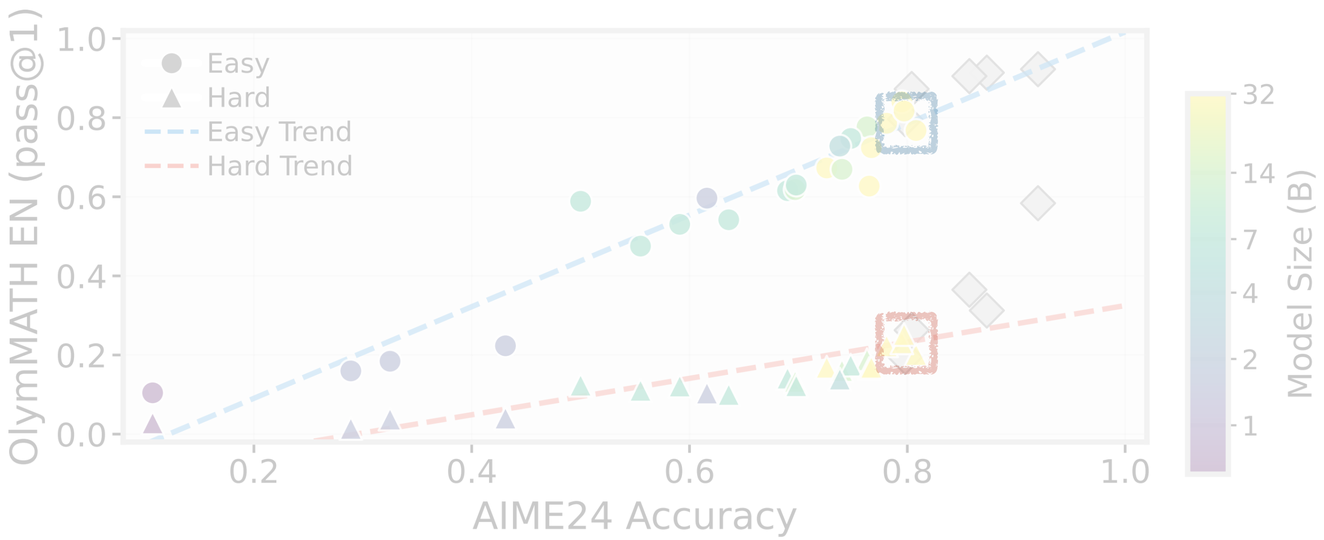
Rigorous Enough?

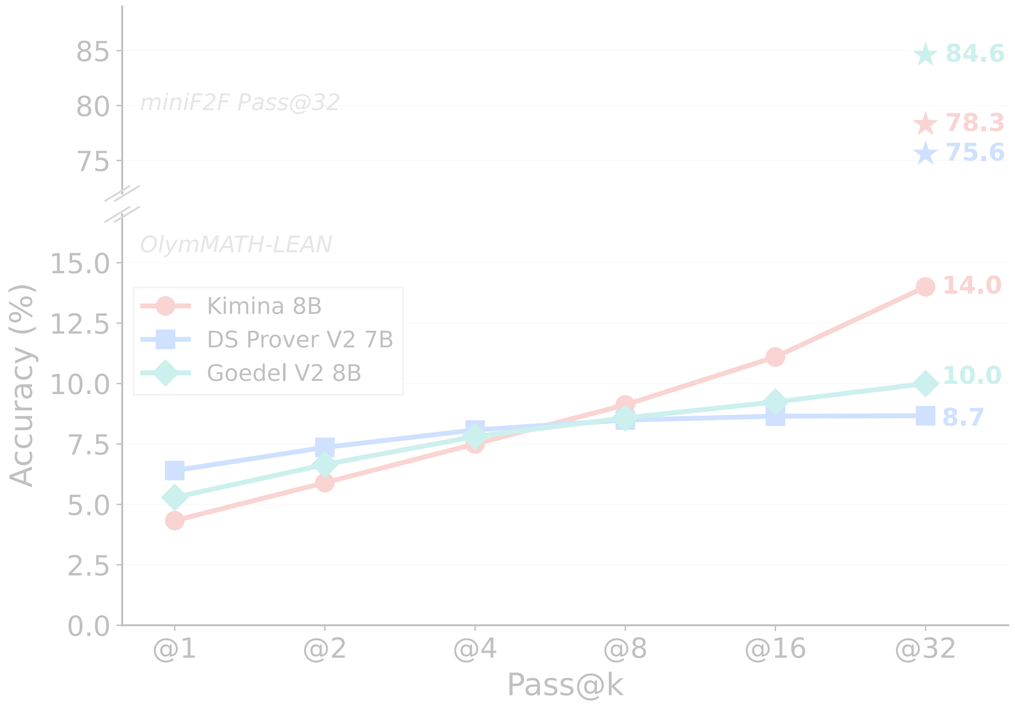




Design

Challenging Enough? Reliable Enough? Rigorous Enough?





Motivation

Challenging Enough? **Reliable Enough?** Rigorous Enough?

Motivation

Challenging Enough? | Reliable Enough? | Rigorous Enough?

- Data Leakage
- Multilingual
- Statistical Stability

Motivation

Challenging Enough? **Reliable Enough?** Rigorous Enough?

- **Data Leakage**
- Multilingual
- Statistical Stability

Motivation

Challenging Enough? **Reliable Enough?** Rigorous Enough?

- **Data Leakage**
- Multilingual
- Statistical Stability

	Test Size	Pile train	DCLM baseline	CC 2025-05	CC 2025-08	CC 2025-13	CC 2025-18	CC 2025-21	CC 2025-26
MMLU	1000	13.20	28.40	13.50	9.00	12.10	11.50	11.70	9.20
MMLU-Pro	1000	5.50	16.20	7.10	5.40	6.00	6.30	7.40	6.90
AIME-2024	30	0.00	0.00	10.00	3.30	6.70	40.00	40.00	13.30
GSM8K	1000	0.00	5.00	5.00	0.80	6.90	0.70	74.20	7.30
MATH-500	500	0.60	3.20	0.60	7.80	0.80	0.80	0.80	8.20
MGSM	250	0.00	0.00	5.60	1.60	35.60	0.80	72.80	6.00

Motivation

Challenging Enough? **Reliable Enough?** Rigorous Enough?

- **Data Leakage**
- Multilingual
- Statistical Stability

	Test Size	Pile train	DCLM baseline	CC 2025-05	CC 2025-08	CC 2025-13	CC 2025-18	CC 2025-21	CC 2025-26
MMLU	1000	13.20	28.40	13.50	9.00	12.10	11.50	11.70	9.20
MMLU-Pro	1000	5.50	16.20	7.10	5.40	6.00	6.30	7.40	6.90
AIME-2024	30	0.00	0.00	10.00	3.30	6.70	40.00	40.00	13.30
GSM8K	1000	0.00	5.00	5.00	0.80	6.90	0.70	74.20	7.30
MATH-500	500	0.60	3.20	0.60	7.80	0.80	0.80	0.80	8.20
MGSM	250	0.00	0.00	5.60	1.60	35.60	0.80	72.80	6.00

Motivation

Challenging Enough? **Reliable Enough?** Rigorous Enough?

- **Data Leakage**
- Multilingual
- Statistical Stability

	Test Size	Pile train	DCLM baseline	CC 2025-05	CC 2025-08	CC 2025-13	CC 2025-18	CC 2025-21	CC 2025-26
MMLU	1000	13.20	28.40	13.50	9.00	12.10	11.50	11.70	9.20
MMLU-Pro	1000	5.50	16.20	7.10	5.40	6.00	6.30	7.40	6.90
AIME-2024	30	0.00	0.00	10.00	3.30	6.70	40.00	40.00	13.30
GSM8K	1000	0.00	5.00	5.00	0.80	6.90	0.70	74.20	7.30
MATH-500	500	0.60	3.20	0.60	7.80	0.80	0.80	0.80	8.20
MGSM	250	0.00	0.00	5.60	1.60	35.60	0.80	72.80	6.00

Example from MATH-500

Problem: For how many positive integers $n > 1$ is it true that 2^{24} is a perfect n^{th} power?
Answer: $\boxed{7}$

Partial Prompt

Prompt: For how many positive integers $n > 1$ is

Qwen2.5-7B Output

it true that 2^{24} is a perfect n^{th} power?
To determine for how many positive integers $n > 1$ the number 2^{24} is a perfect n-th power...
Let's count these values. There are 7 such values...
The final answer is $\boxed{7}$.

Llama3.1-8B Output

the set of positive divisors of n (including 1 and n) less than 50 ?
(A) 15 **(B)** 16 **(C)** 17 ...
The number of positive divisors of n is given by...
The number of factors of each of these numbers is...
The answer is $\boxed{\textbf{(B)}\ 16}$

Motivation

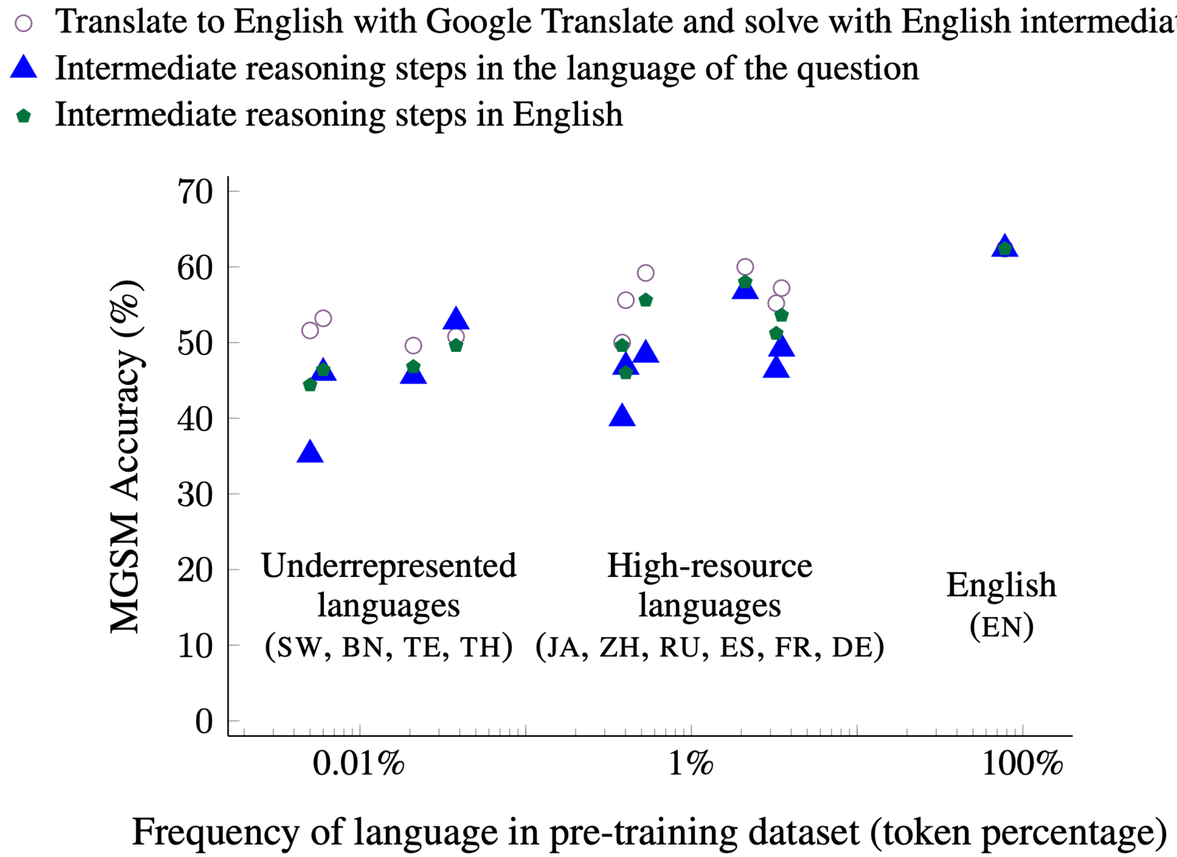
Challenging Enough? | Reliable Enough? | Rigorous Enough?

- Data Leakage
- **Multilingual**
- Statistical Stability

Motivation

Challenging Enough? **Reliable Enough?** Rigorous Enough?

- Data Leakage
- **Multilingual**
- Statistical Stability



Motivation

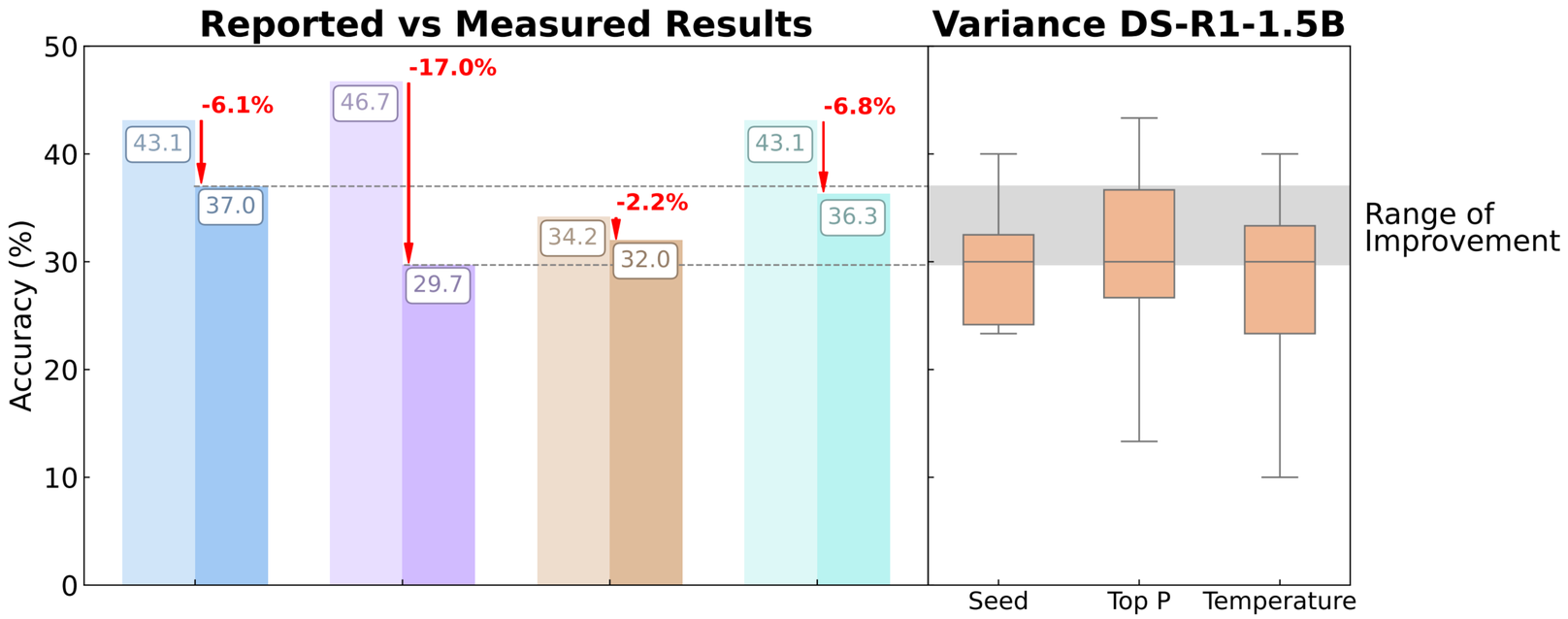
Challenging Enough? **Reliable Enough?** Rigorous Enough?

- Data Leakage
- Multilingual
- **Statistical Stability**

Motivation

Challenging Enough? **Reliable Enough?** Rigorous Enough?

- Data Leakage
- Multilingual
- **Statistical Stability**



Design

Challenging Enough? | **Reliable Enough?** | Rigorous Enough?

- Data Leakage
- Multilingual
- Statistical Stability

Design

Challenging Enough? | Reliable Enough? | Rigorous Enough?

- Data Leakage
- Multilingual
- Statistical Stability

Design

Challenging Enough? **Reliable Enough?** Rigorous Enough?

- **Data Leakage**
  - We collected all problems from printed materials w/ LLM quick verification
  - **Informal:** Human rewrite + verify **Formal:** Multi-agent system w/ Lean Server
- **Multilingual**
- **Statistical Stability**

Design

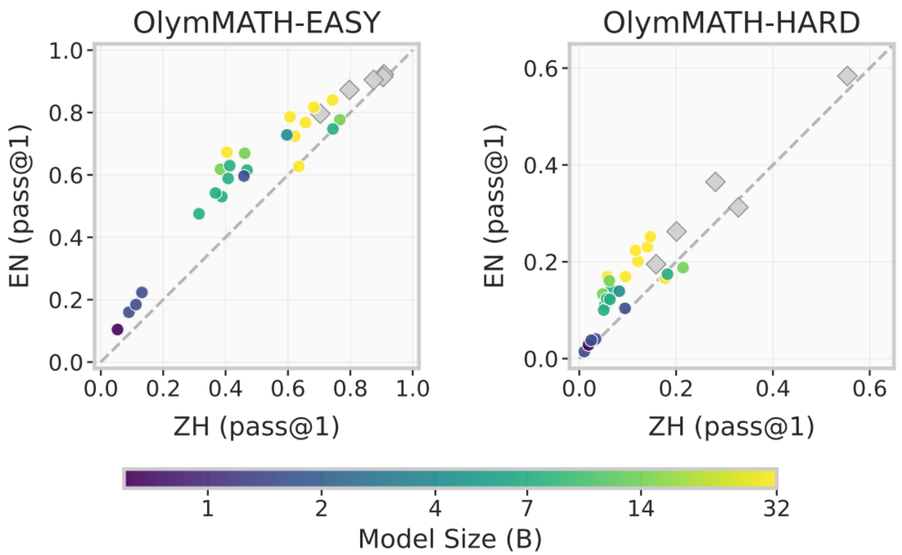
Challenging Enough? **Reliable Enough?** Rigorous Enough?

- **Data Leakage**
  - We collected all problems from printed materials w/ LLM quick verification
  - **Informal:** Human rewrite + verify **Formal:** Multi-agent system w/ Lean Server
- **Multilingual**
  - Problems are translated using Claude Sonnet 3.7 (translate) + GPT 4o (verify)
  - After convergence, human do the final check
- **Statistical Stability**

Design

Challenging Enough? **Reliable Enough?** Rigorous Enough?

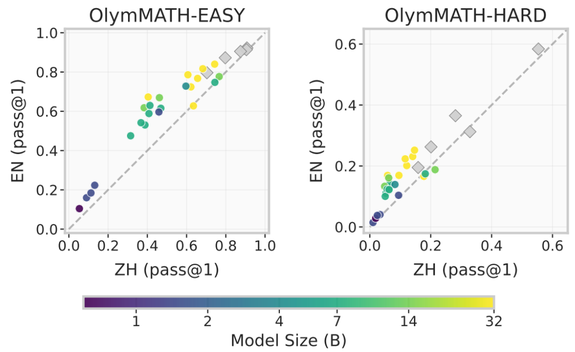
- **Data Leakage**
  - We collected all problems from printed materials w/ LLM quick verification
  - **Informal:** Human rewrite + verify **Formal:** Multi-agent system w/ Lean Server
  - Problems are translated using Claude Sonnet 3.7 (translate) + GPT 4o (verify)
  - After convergence, human do the final check
- **Multilingual**
- **Statistical Stability**



Design

Challenging Enough? **Reliable Enough?** Rigorous Enough?

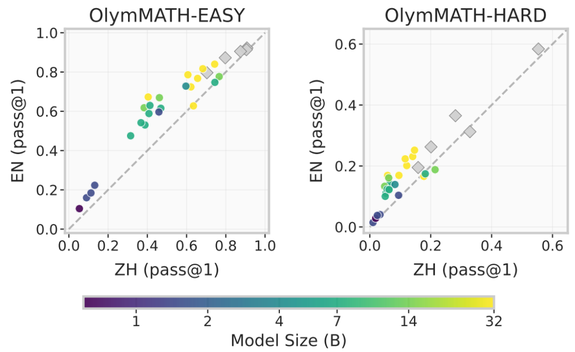
- **Data Leakage**
  - We collected all problems from printed materials w/ LLM quick verification
  - **Informal:** Human rewrite + verify **Formal:** Multi-agent system w/ Lean Server
- **Multilingual**
  - Problems are translated using Claude Sonnet 3.7 (translate) + GPT 4o (verify)
  - After convergence, human do the final check
- **Statistical Stability**



Design

Challenging Enough? **Reliable Enough?** Rigorous Enough?

- **Data Leakage**
  - We collected all problems from printed materials w/ LLM quick verification
  - **Informal:** Human rewrite + verify **Formal:** Multi-agent system w/ Lean Server
  - Problems are translated using Claude Sonnet 3.7 (translate) + GPT 4o (verify)
  - After convergence, human do the final check
- **Multilingual**



- **Statistical Stability**
  - We have significant more problems than concurrent works (w/ same difficulty)

Motivation

Challenging Enough?

Reliable Enough?

Rigorous Enough?

Motivation

Challenging Enough? Reliable Enough? Rigorous Enough?

Problem: In the triangle $\triangle ABC$, the lengths of the three sides are a, b, c, satisfying the condition $2a^2 + b^2 + c^2 = 4$. Find the maximum value of the area of $\triangle ABC$.

Answer: $\frac{\sqrt{5}}{5}$. (When $a = \frac{2}{\sqrt{5}}, b = c = \sqrt{\frac{6}{5}}$)

Motivation

Challenging Enough? Reliable Enough? Rigorous Enough?

Problem: In the triangle $\triangle ABC$, the lengths of the three sides are a, b, c, satisfying the condition $2a^2 + b^2 + c^2 = 4$. Find the maximum value of the area of $\triangle ABC$.

Answer: $\frac{\sqrt{5}}{5}$. (When $a = \frac{2}{\sqrt{5}}, b = c = \sqrt{\frac{6}{5}}$)

Model Output: We wish to maximize the area of $\triangle ABC$, subject to $2a^2 + b^2 + c^2 = 4$. Because the condition is symmetric in b and c (but not in a), it is reasonable to look for an optimum among triangles having $b = c$. (Later one can check that assuming $b = c$ indeed gives the maximum.)

The problem involves maximizing the area of an isosceles triangle with sides a, b, b under the constraint $2a^2 + b^2 + c^2 = 4$. By assuming $b = c$, we derive a formula for the area in terms of b and find the maximum area by expressing it as a function of $u = b^2$. We then determine the critical points of this quadratic function, identify the maximum area, and finally recover the side lengths to verify that they satisfy the original constraints. The maximum area achieved is $\frac{\sqrt{5}}{5}$.

Motivation

Challenging Enough? Reliable Enough? Rigorous Enough?

Problem: In the triangle $\triangle ABC$, the lengths of the three sides are a, b, c, satisfying the condition $2a^2 + b^2 + c^2 = 4$. Find the maximum value of the area of $\triangle ABC$.

Answer: $\frac{\sqrt{5}}{5}$. (When $a = \frac{2}{\sqrt{5}}, b = c = \sqrt{\frac{6}{5}}$)

Model Output: We wish to maximize the area of $\triangle ABC$, subject to $2a^2 + b^2 + c^2 = 4$. Because the condition is symmetric in b and c (but not in a), it is reasonable to look for an optimum among triangles having $b = c$. (Later one can check that assuming $b = c$ indeed gives the maximum.)

The problem involves maximizing the area of an isosceles triangle with sides a, b, b under the constraint $2a^2 + b^2 + c^2 = 4$. By assuming $b = c$, we derive a formula for the area in terms of b and find the maximum area by expressing it as a function of $u = b^2$. We then determine the critical points of this quadratic function, identify the maximum area, and finally recover the side lengths to verify that they satisfy the original constraints. The maximum area achieved is $\frac{\sqrt{5}}{5}$.

Design

Challenging Enough?

Reliable Enough?

Rigorous Enough?

Design

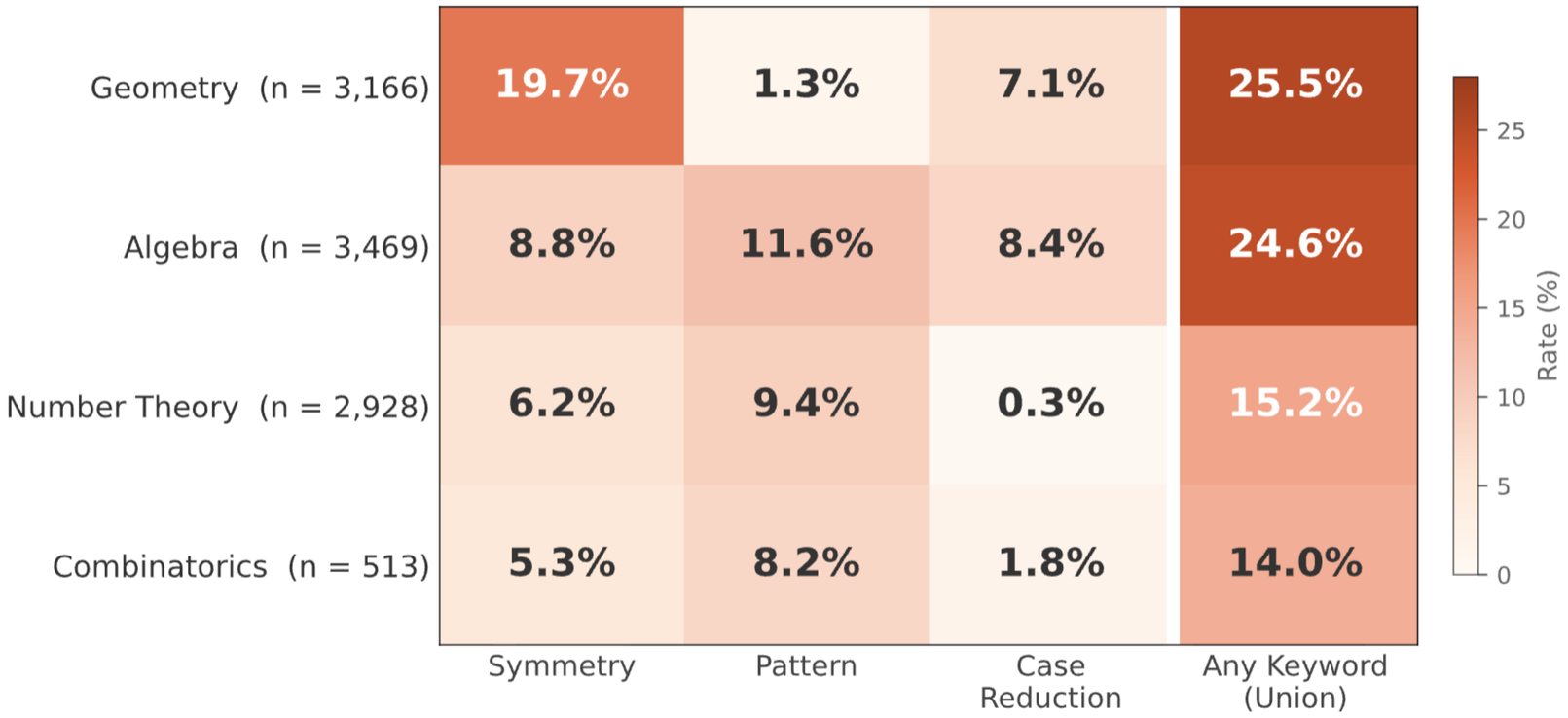
Challenging Enough? Reliable Enough? Rigorous Enough?

Let's detect the keywords! (WLOG, etc.)

Design

Challenging Enough? Reliable Enough? **Rigorous Enough?**

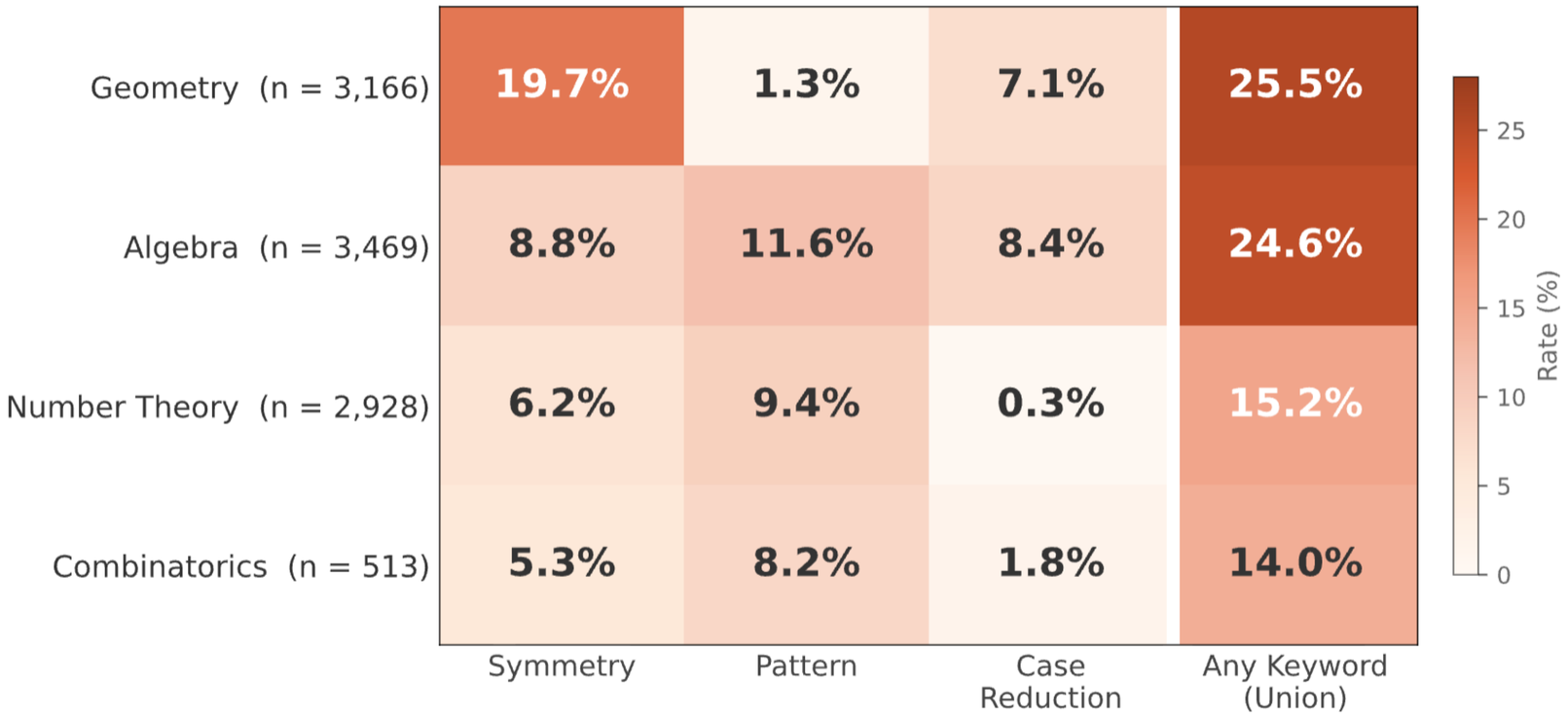
Let's detect the keywords! (WLOG, etc.)



Design

Challenging Enough? Reliable Enough? **Rigorous Enough?**

Let's detect the keywords! (WLOG, etc.)



Stronger models even have higher keyword rate!

Design

Challenging Enough?

Reliable Enough?

Rigorous Enough?

Design

Challenging Enough? Reliable Enough? Rigorous Enough?

Let's do LLM-as-Judge?

Design

Challenging Enough? Reliable Enough? **Rigorous Enough?**

Let's do LLM-as-Judge?

LLM Evaluators Recognize and Favor Their Own Generations, …

Design

Challenging Enough? Reliable Enough? Rigorous Enough?

Let's do LLM-as-Judge?

Let's create a Lean subset!

Design

Challenging Enough? Reliable Enough? Rigorous Enough?

Let's do LLM-as-Judge?

Let's create a Lean subset!

Lean encodes proofs as programs, so proof correctness reduces to type-checking.

Design

Challenging Enough? Reliable Enough? **Rigorous Enough?**

Let's do LLM-as-Judge?

Let's create a Lean subset!

Lean encodes proofs as programs, so proof correctness reduces to type-checking.

- **Clean Up & Verification**: Claude Opus 4.5 is used to clean up LaTeX syntax and format. 3 DeepSeek V3.2 Speciale are used to check the proof from collected material.
- **Iterative Refine w/ environment**: Claude Opus 4.5 is prompted to interact with Lean REPL server and iteratively refine (Can be parallelized using Claude Code subagent)
- **Final Check**: 3 Gemini 3.0 Flash are prompted to verify

Design

Challenging Enough? Reliable Enough? **Rigorous Enough?**

Let's do LLM-as-Judge?

Let's create a Lean subset!

Lean encodes proofs as programs, so proof correctness reduces to type-checking.

- **Clean Up & Verification**: Claude Opus 4.5 is used to clean up LaTeX syntax and format. 3 DeepSeek V3.2 Speciale are used to check the proof from collected material.
- **Iterative Refine w/ environment**: Claude Opus 4.5 is prompted to interact with Lean REPL server and iteratively refine (Can be parallelized using Claude Code subagent)
- **Final Check**: 3 Gemini 3.0 Flash are prompted to verify

Please see more details in our paper!

Conclusion:

Conclusion: OlymMATH

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

Subject: Number Theory. **Formal Statement:**

```
theorem to_prove
  (n : Nat) (p : Nat) (hp : p.Prime)
  (hdiv : p | 2^n + 1) : p % 8 != 7 := by sorry
```

OlymMATH-EASY

100 Problems

AIME Level

EN / CH

OlymMATH-HARD

100 Problems

More Challenging

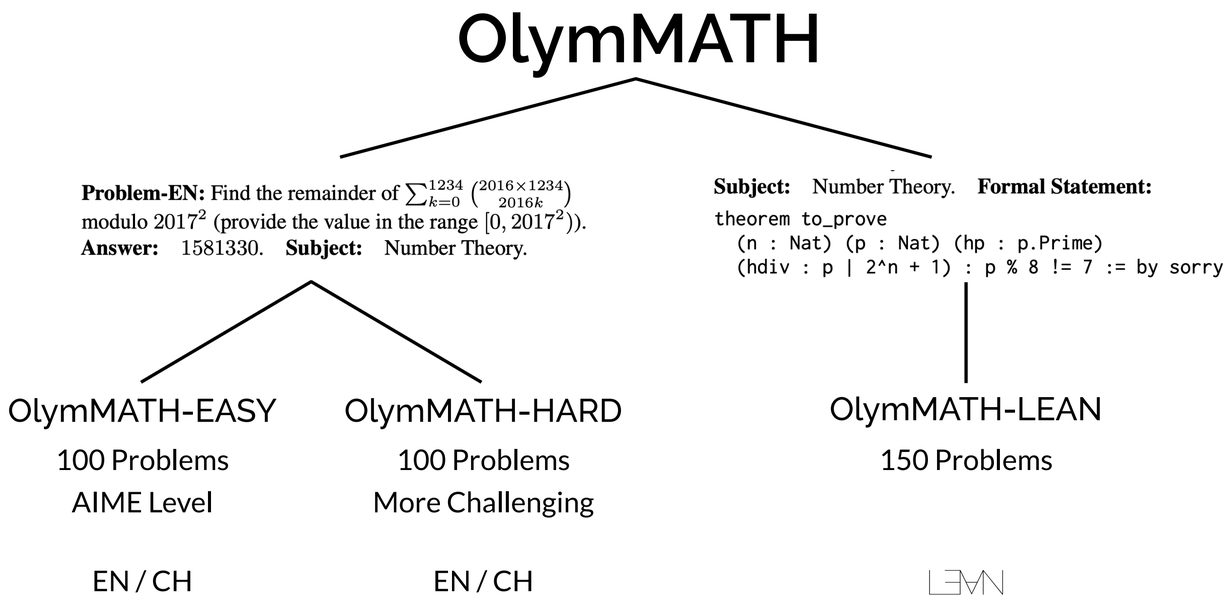
EN / CH

OlymMATH-LEAN

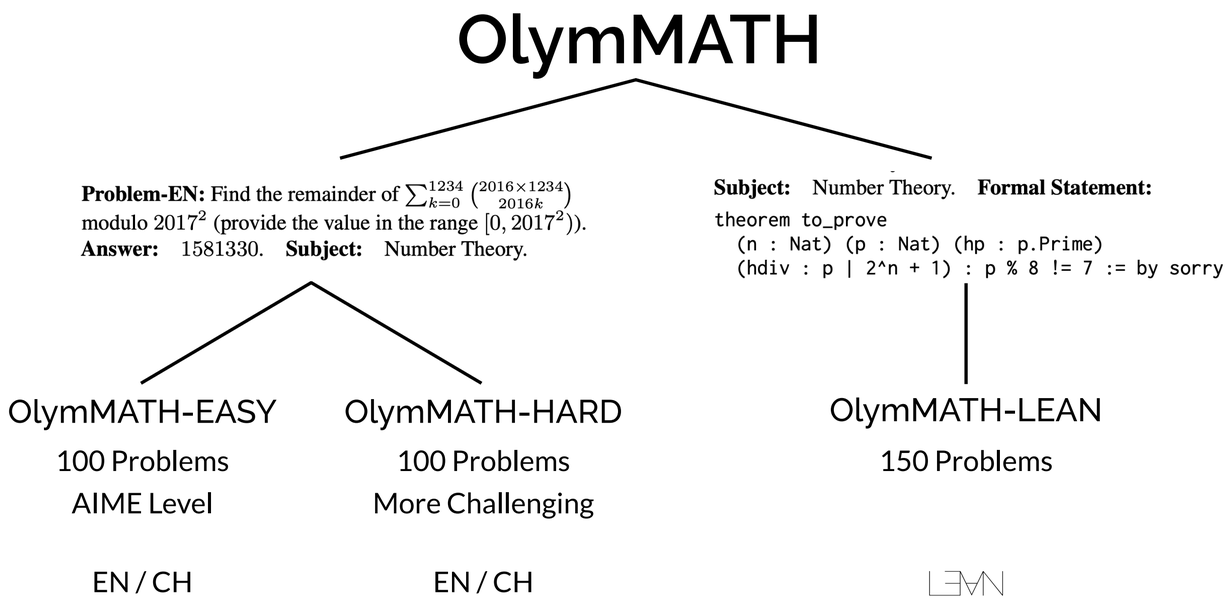
150 Problems

LEAN

Conclusion:


OlymMATH
Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016\times1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. Subject: Number Theory.
Subject: Number Theory. Formal Statement:
theorem to_prove
(n : Nat) (p : Nat) (hp : p.Prime)
(hdiv : p | 2^n + 1) : p % 8 != 7 := by sorry
OlymMATH-EASY
100 Problems
AIME Level
EN / CH
OlymMATH-HARD
100 Problems
More Challenging
EN / CH
OlymMATH-LEAN
150 Problems

Conclusion:



- **Challenging**: Our EASY subset aligns AIME difficulty (to provide more robust evaluation of weaker LLMs); HARD and LEAN are even more challenging
- **Reliable**: Expert manual verification; Bilingual evaluation reveals EN-ZH gap; More statistically stable evaluation
- **Rigorous**: More fine-grained process-level verfication with Lean
- **Open-source**: Releasing 582k+ reasoning trajectories, a visualization tool, expert solutions, and Lean synthesis multi-agent pipeline

OlymMATH in the Wild

- **RL Training & Optimization**: Pass@k Training, Beyond Pass@1 (SvS), TreeAdv, R^2VPO, GeoRA, Adaptive Ability Decomposing (A^2D), …
- **Distillation & Data Selection**: Agent Distillation, From Correction to Mastery (SCoRe), Unified Data Selection (HES), …
- **Reasoning & Tool Use**: Code-Integrated Reasoning (CIR), Sticker-TTS, Agentic RL Scaling Law (ZeroTIR), …
- **Multi-agent Systems**: AgentDropoutV2, HSM3K-CN, …

Towards Better Evaluation of LLMs

- **Contamination Resistance**: Live & Dynamic Benchmarks

  MathArena SciDA MathDuels

- **Extended Difficulty**: Scaling from competition-level to research-level

  REALMATH CompMath-MCQ Proof2Hybrid Re^{2}Math

- **Beyond Final-answer Accuracy**: Evaluating diversity, proof rigor, …

  RIMO PROOFGRID Think360°

- **Multilingual & Multimodal**: Evaluating more languages & modalities

  MATHNET IRLBench GEOGRAMBENCH OMIBench

- **More Domains**: Evaluating reasoning on more domains

  PHYBench CMedCalc-Bench ModelingBench HumorBench

Thanks for listening!

Paper
https://arxiv.org/abs/2503.21380

GitHub
https://github.com/RUCAIBox/OlymMATH

Hugging Face
https://huggingface.co/datasets/RUC-AIBOX/OlymMATH

Have any questions? Please send an email to `hxiang.sun@gmail.com`

X: @CoderBak Personal Website: `hxiang-sun.github.io`