Stanford HAI Stanford University Human-Centered Artificial Intelligence

Stanford Artificial Intelligence

Berkeley UNIVERSITY OF CALIFORNIA

Massachusetts Institute of Technology

Solving Inequality Proofs with Large Language Models

Jiayi Sheng*[1,2], Luna Lyu*[1], Jikai Jin[1], Tony Xia[1], Alex Gu[3], James Zou†[1], Pan Lu†[1]

[1] Stanford University [2] UC Berkeley [3] Massachusetts Institute of Technology

* Co–first authors † Co–senior authors

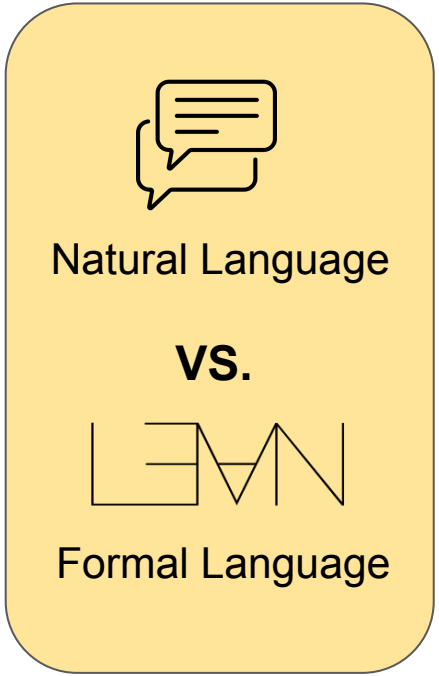
NeurIPS 2025 Spotlight

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

- Most existing inequality benchmarks are represented in **formal language** (such as Lean).
- Informal reasoning is closer to **human intuition**.
- LLM trained on natural language corpora has **potential informal inequality solving ability**.

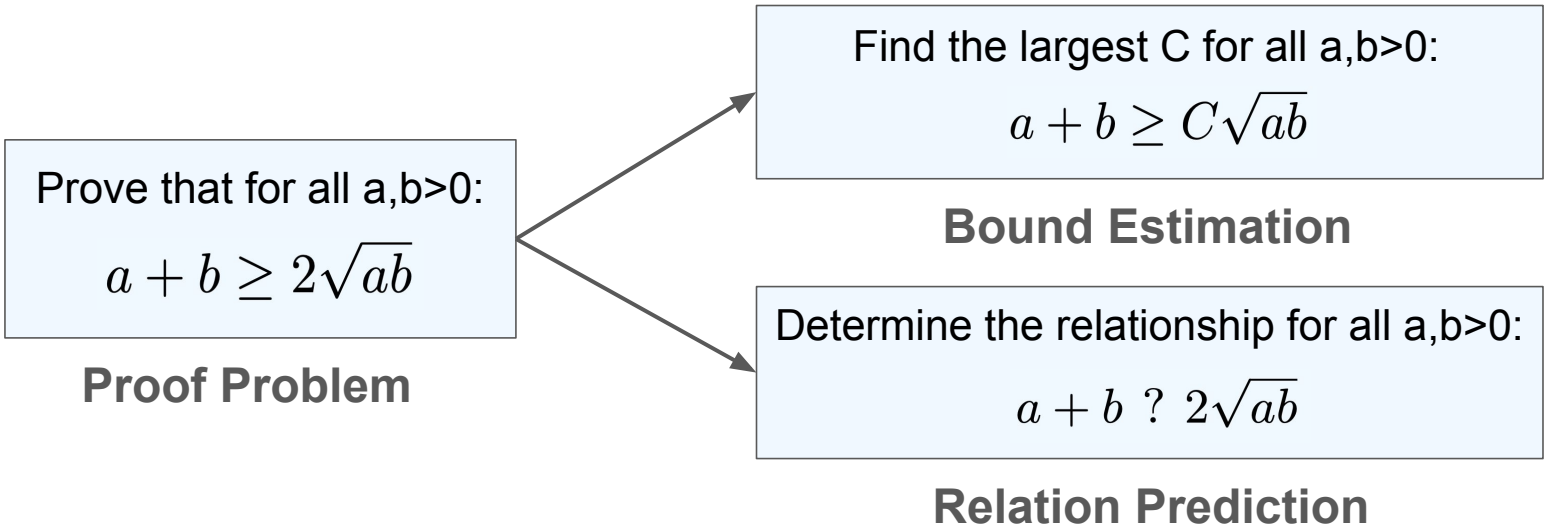


Goal: Evaluate LLMs on informal inequality proof problems.

Task Reformulation

We reformulate the inequality proofs in to two **informal** yet **verifiable** subtasks:

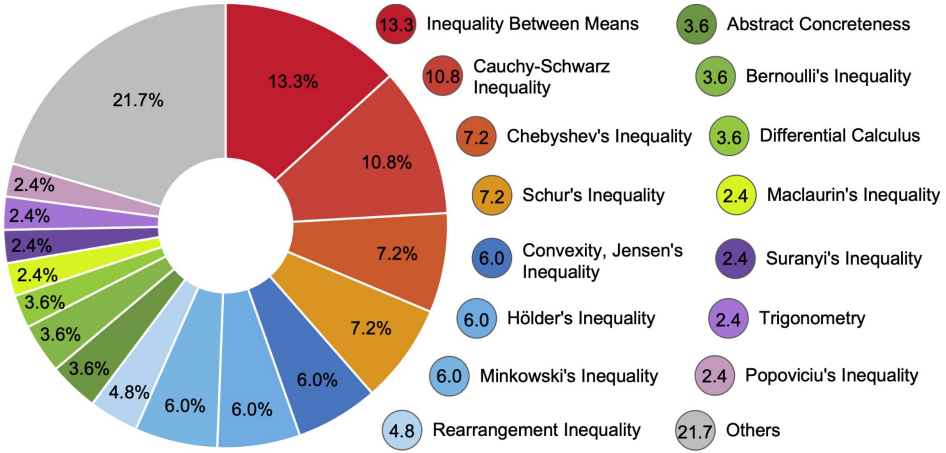
- **Bound Estimation**: Finding an optimal constant for a given inequality.
- **Relation Prediction**: Determining the relationship between two expressions.



IneqMath Dataset

- Each training problem includes **up to four step-wise solutions**.
- 76.8% are annotated with relevant **theorems**.
- Test problems are crafted by **IMO medalists** to ensure difficulty.

Statistic	Number	Bnd.	Rel.
Theorem categories	29	-	-
Named theorems	83	-	-
Training problems (for training)	1252	626	626
- With theorem annotations	962	482	480
- With solution annotations	1252	626	626
- Avg. solutions per problem	1.05	1.06	1.05
- Max solutions per problem	4	4	4
Dev problems (for development)	100	50	50
Test problems (for benchmarking)	200	96	104



IneqMath Training Examples

NEURAL INFORMATION PROCESSING SYSTEMS

IneqMath Training Example 1: Bound Problem

Question: Find the maximal constant C such that for all real numbers a, b, c, the inequality holds:

$$\sqrt{a^2+(1-b)^2}+\sqrt{b^2+(1-c)^2}+\sqrt{c^2+(1-a)^2} \geq C$$

Solution: Applying Minkowsky's Inequality to the left-hand side we have

$$\sqrt{a^2+(1-b)^2}+\sqrt{b^2+(1-c)^2}+\sqrt{c^2+(1-a)^2} \geq \sqrt{(a+b+c)^2+(3-a-b-c)^2}$$

By denoting $a+b+c=x$, we get

$$\sqrt{(a+b+c)^2+(3-a-b-c)^2}=\sqrt{2\left(x-\frac{3}{2}\right)^2+\frac{9}{2}} \geq \sqrt{\frac{9}{2}}=\boxed{\frac{3\sqrt{2}}{2}}.$$

Minkowsky's Inequality Theorem: For any real number $r \geq 1$ and any positive real numbers $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$

$$\left(\sum_{i=1}^{n}(a_i+b_i)^r\right)^{\frac{1}{r}} \leq \left(\sum_{i=1}^{n}a_i^r\right)^{\frac{1}{r}}+\left(\sum_{i=1}^{n}b_i^r\right)^{\frac{1}{r}}$$

IneqMath Testing Examples

NEURAL INFORMATION PROCESSING SYSTEMS

IneqMath Testing Example 1: Bound Problem

Question: Let $a_1, a_2, \ldots, a_n > 0$ such that $a_1 + a_2 + \ldots + a_n < 1$. Determine the minimal constant C such that the following inequality holds for all $a_1, a_2, \ldots, a_n$:

$$\frac{a_1 \cdot a_2 \ldots a_n \left(1 - a_1 - a_2 - \ldots - a_n\right)}{\left(a_1 + a_2 + \ldots + a_n\right)\left(1 - a_1\right)\left(1 - a_2\right)\ldots\left(1 - a_n\right)} \leq C \frac{3}{n^{n-1}}.$$

IneqMath Testing Example 2: Relation Problem

Question: Let a, b, c be the sides of any triangle. Consider the following inequality:

$$3\left(\sum_{cyc} ab\left(1 + 2\cos(c)\right)\right) \quad (\quad) \quad 2\left(\sum_{cyc} \sqrt{\left(c^2 + ab(1 + 2\cos(c))\right)\left(b^2 + ac(1 + \cos(b))\right)}\right).$$

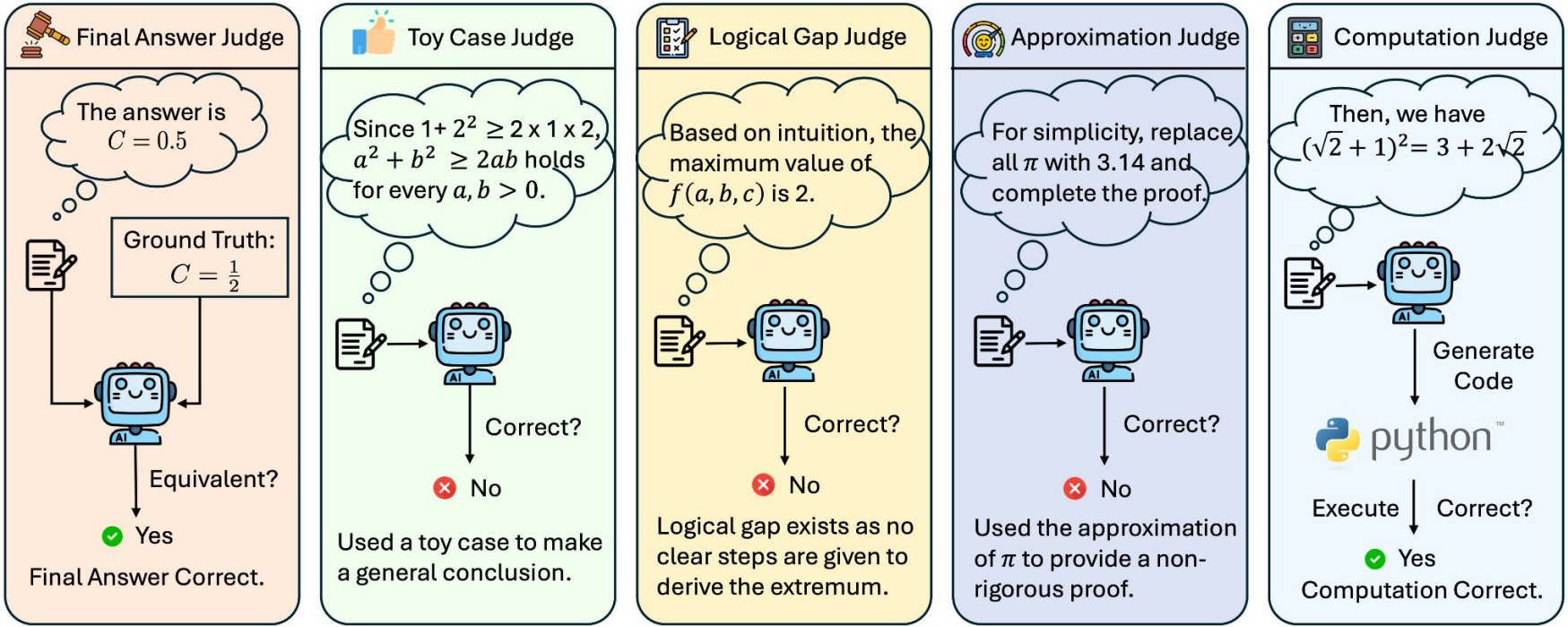
Determine the correct inequality relation to fill in the blank.

Options: (A) $\leq$ (B) $\geq$ (C) $=$ (D) $<$ (E) $>$ (F) None of the above

Fine-grained Informal Judges

NEURAL INFORMATION PROCESSING SYSTEMS

- One **Final Answer** Judge + four **Step-wise** Judges.



Judges Performance

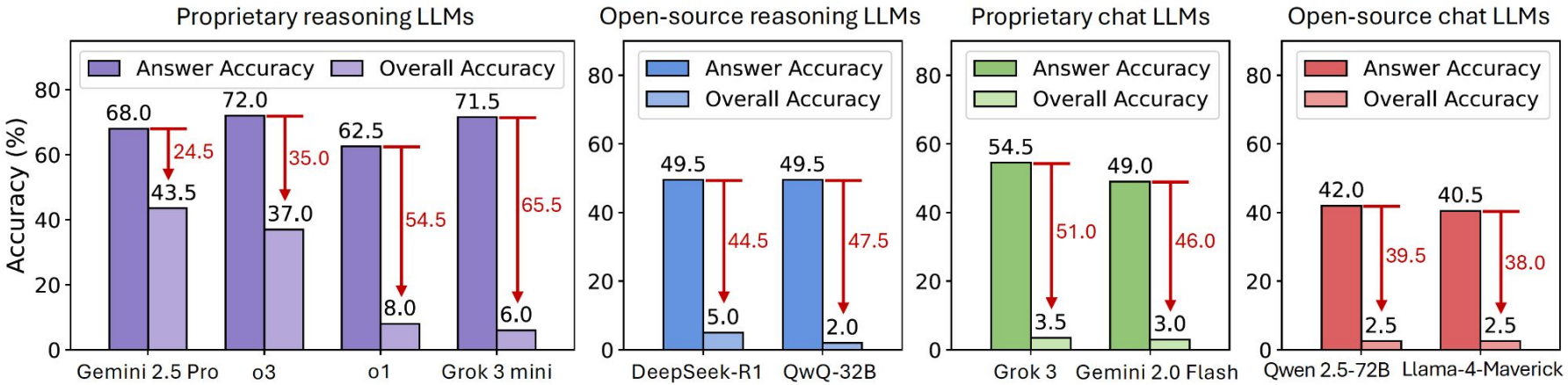
- We evaluate the informal judges on the dev set with human annotations.
- The judges achieve **strong alignment** with human annotations with **F1 = 0.93**.

LLM-as-Judge	Judge type	Accuracy	Precision	Recall	F1 score
Final Answer Judge	Answer checking	1.00	1.00	1.00	1.00
Toy Case Judge	Step soundness	0.91	0.86	0.97	0.91
Logical Gap Judge	Step soundness	0.96	0.95	0.98	0.96
Numerical Approximation Judge	Step soundness	0.96	0.95	0.98	0.96
Numerical Computation Judge	Step soundness	0.71	0.68	0.98	0.80
Average	-	0.91	0.89	0.98	**0.93**

Key Findings 1 — Soundness Gap Exists

NEURAL INFORMATION PROCESSING SYSTEMS

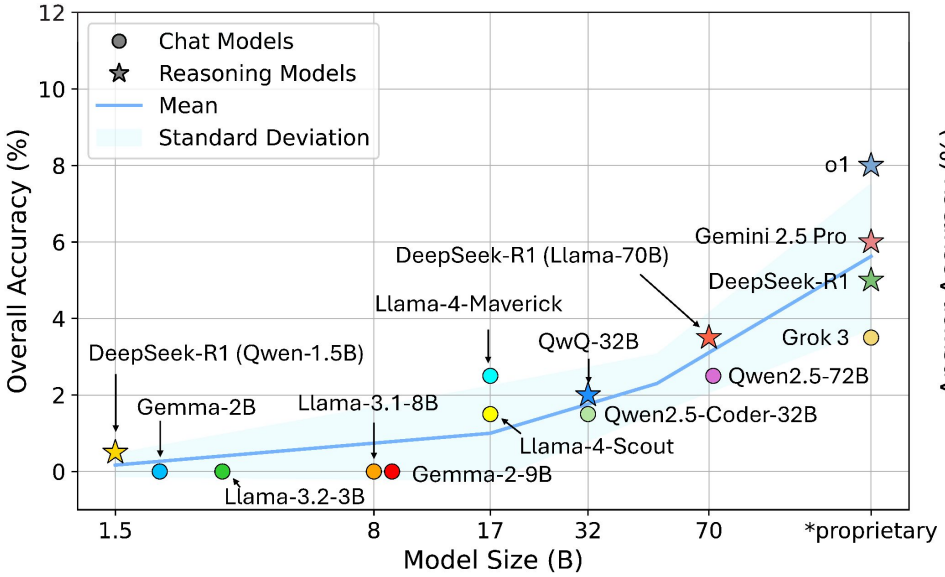
- **Overall Accuracy**: Correct final answer + ALL reasoning steps sound.
- **Answer Accuracy:** Correct final answer, how it got there doesn't matter.
- **Findings**: LLMs often **guess the right answer** for complex Olympiad-level inequalities, but their **step-by-step reasoning is unsound**.

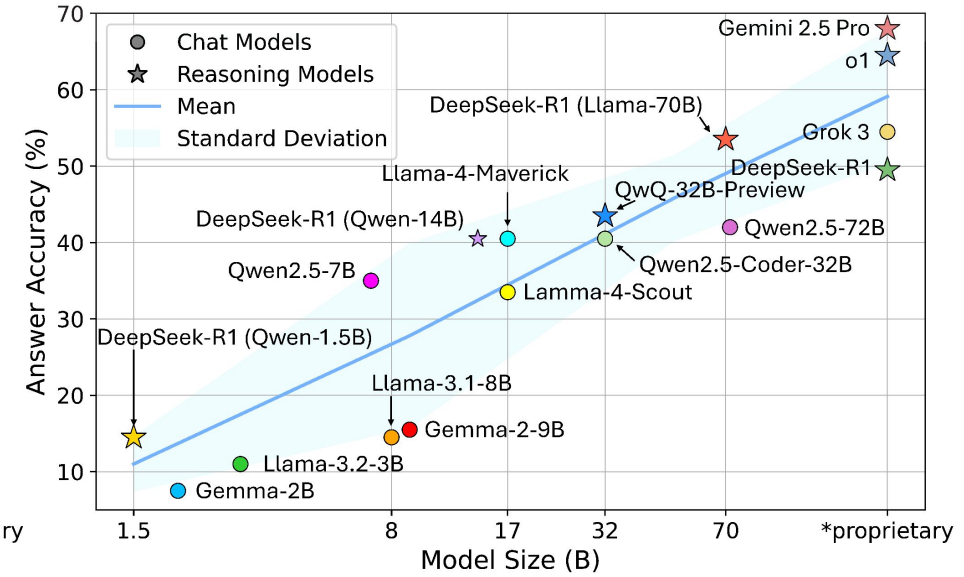


Key Findings 2 — Bigger Isn't Always Better

NEURAL INFORMATION PROCESSING SYSTEMS

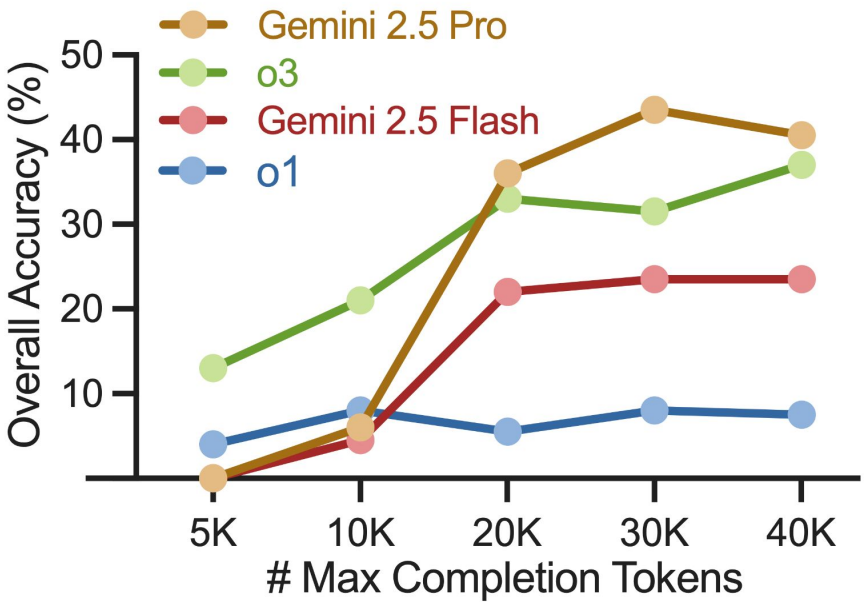
- Bigger models get **more final answers right**. (Right)
- But the increased model size is **insufficient to enhance overall accuracy**. (Left)





Key Findings 3 — Thinking Longer Can't Solve the Soundness Gap

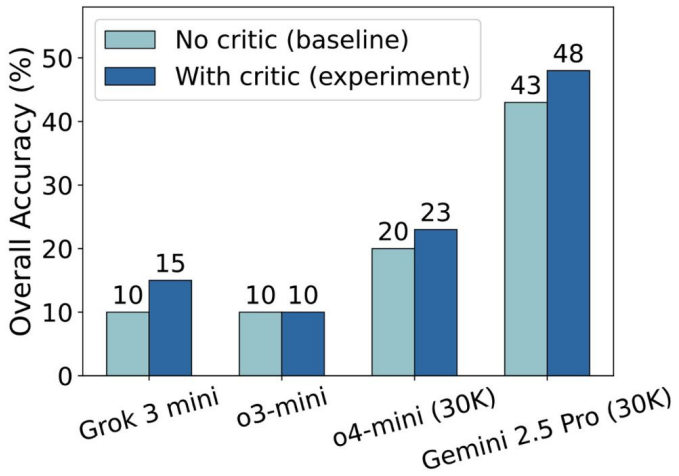
- Performance gains tend to **saturate** as the maximum token limit increases.
- Simply extending the reasoning chain offers **diminishing returns** for overall proof correctness.

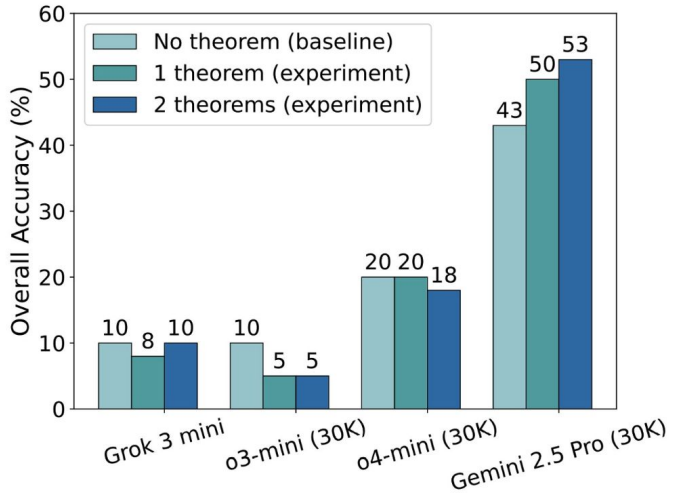


Improvement Strategies

NEURAL INFORMATION PROCESSING SYSTEMS

- **Self-improvement (critic-guided)**: Gemini 2.5 Pro's overall accuracy up **+5%** (43%→48%) via self-critique! (Left)
- **Theorem augmentation (providing key theorem hints)**: Gemini 2.5 Pro's overall accuracy up another **+10%** with theorem guidance! (Right)





Stanford HAI Stanford University Human-Centered Artificial Intelligence

Stanford Artificial Intelligence

Berkeley University of California

Massachusetts Institute of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

Come visit our poster!

Wed 3 Dec 4:30 p.m. PST — 7:30 p.m. PST

IneqMath

https://ineqmath.github.io/

Project Website

Email: lupantech@gmail.com, jamesz@stanford.edu, jiayi_sheng@berkeley.edu