Solving Inequality Proofs with Large Language Models

Stanford ARTIFICIAL INTELLIGENCE

Jiayi Sheng*[β], Luna Lyu*[α], Jikai Jin[α], Tony Xia[α], Alex Gu[γ], James Zou[†α], Pan Lu[†α]

[α] Stanford University [β] UC Berkeley [γ] MIT

* Co-first authors † Co-senior authors

IneqMath

Project Website

https://ineqmath.github.io/

Introduction

Mathematics demands rigorous proofs, not just correct answers. This is especially crucial for inequality problems. **Do large language models truly understand proofs, or just guess correct answers?**

To explore this question, we introduce:

1. **A novel reformulation:** Decomposing inequality proving into **informal**, **verifiable** subtasks (bound estimation & relation prediction).

2. **IneqMath:** An expert-curated benchmark of Olympiad inequalities + step-by-step solution & **theorem**.

3. **LLM-as-Judge**: A framework for rigorously evaluating both final answers AND step-by-step soundness.

Task Reformulation

To bridge formal verification with natural language, we reformulate inequality proofs into two **informal yet verifiable** subtasks.

Every Inequality problem can be formed as a triple:

$$\Pi = \big(f(\mathbf{x}),\ g(\mathbf{x}),\ \mathcal{D}\big),\ \text{where } f,g:\mathcal{D}\to\mathbb{R},\ \mathcal{D}\subseteq\mathbb{R}^n$$

- Bound estimation: determine the extremal with $g(x) > 0$:

$$C^\star \ =\ \sup\{C : f(\mathbf{x}) \geq Cg(\mathbf{x}), \forall \mathbf{x}\in\mathcal{D}\} \text{ or } \inf\{C : f(\mathbf{x}) \leq Cg(\mathbf{x}), \forall \mathbf{x}\in\mathcal{D}\}$$

- Relation prediction: predict the relationship between $f(x)$ and $g(x)$ for all $x \in \mathcal{D}$ (i.e. $>$, $\geq$, $=$, $\leq$, $<$, or none of the above).

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 Example 2: Relation Problem

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

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

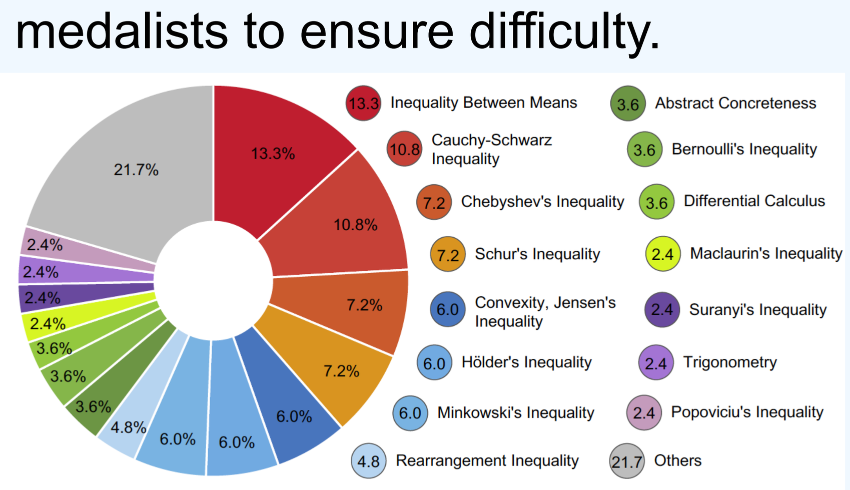
Determine the correct inequality relation to fill in the blank.

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

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.

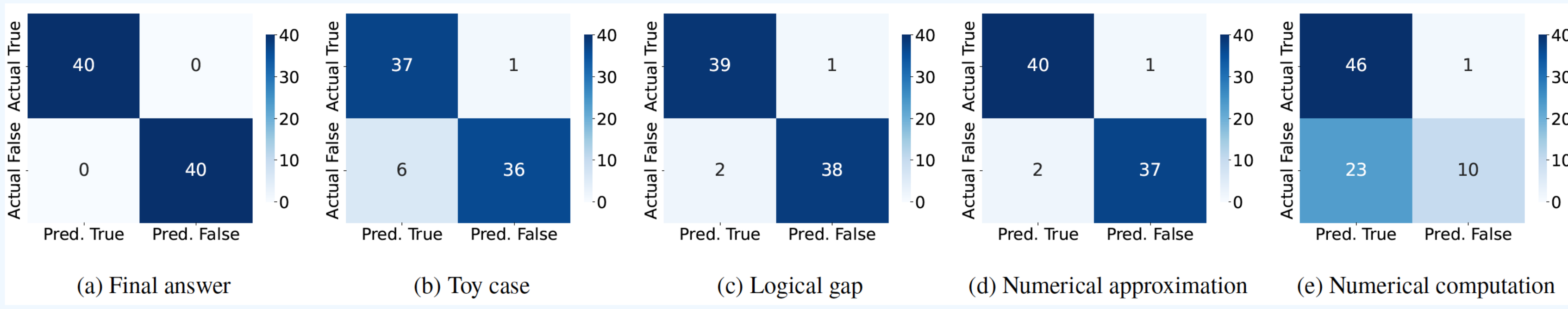
Statistic	Number	Bnd.	Rel.
Theorem categories	29	-	-
Named theorems	83	-	-
Training problems (for training)	1252	626	626
- With theorem annotations	962	482	480
- With solution annotations	1252	626	626
- Avg. solutions per problem	1.05	1.06	1.05
- Max solutions per problem	4	4	4
Dev problems (for development)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



	Data Source		Data Annotation		Problem and Evaluation		
Datasets	Training	Test / Dev	#Theorem	Solution	Category	Format	Evaluation
INT	Synthesized	Synthesized	35	✓	Proof	Formal	Symbolic DSL
AIPS	Synthesized	✗	8	✓	Proof	Formal	Symbolic DSL
MO-INT	✗	Data compilation	✗	✗	Proof	Formal	Symbolic DSL
MINIF2F	✗	Autoformalization	✗	✗	Proof	Formal	LEAN
ProofNet	✗	Autoformalization	✗	✗	Proof	Formal	LEAN
FormalMATH	✗	Autoformalization	✗	✗	Proof	Formal	LEAN
leanWorkbook	Autoformalization	Autoformalization	✗	✗	Proof	Formal	LEAN
Proof or Bluff	✗	Data compilation	✗	✗	Proof	Informal	Human judge
CHAMP	✗	Autoformalization	✗	✗	Open	Informal	Human judge
Putnam Axiom	✗	Data compilation	✗	✗	Open	Informal	Answer checking
LiveMathBench	✗	Data compilation	✗	✗	Open	Informal	Answer checking
IneqMath (Ours)	Expert annotated	Expert annotated	83	✓	MC, Open	Informal	LLM-as-judge

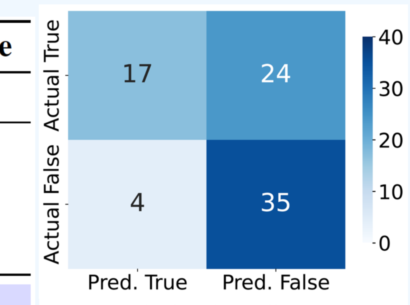
Fine-grained LLM Judges

- **Final Answer Judge**: Validates correctness of the final answer.
- **Toy Case Judge**: Checks for overgeneralization from special cases.
- **Logical Gap Judge**: Detects skipped logical steps.
- **Numerical Approximation Judge**: Flags inappropriate approximations.
- **Numerical Computation Judge**: Identifies arithmetic errors.
- A solution is deemed correct overall only if it passes all five judges.



(a) Final answer (b) Toy case (c) Logical gap (d) Numerical approximation (e) Numerical computation

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**

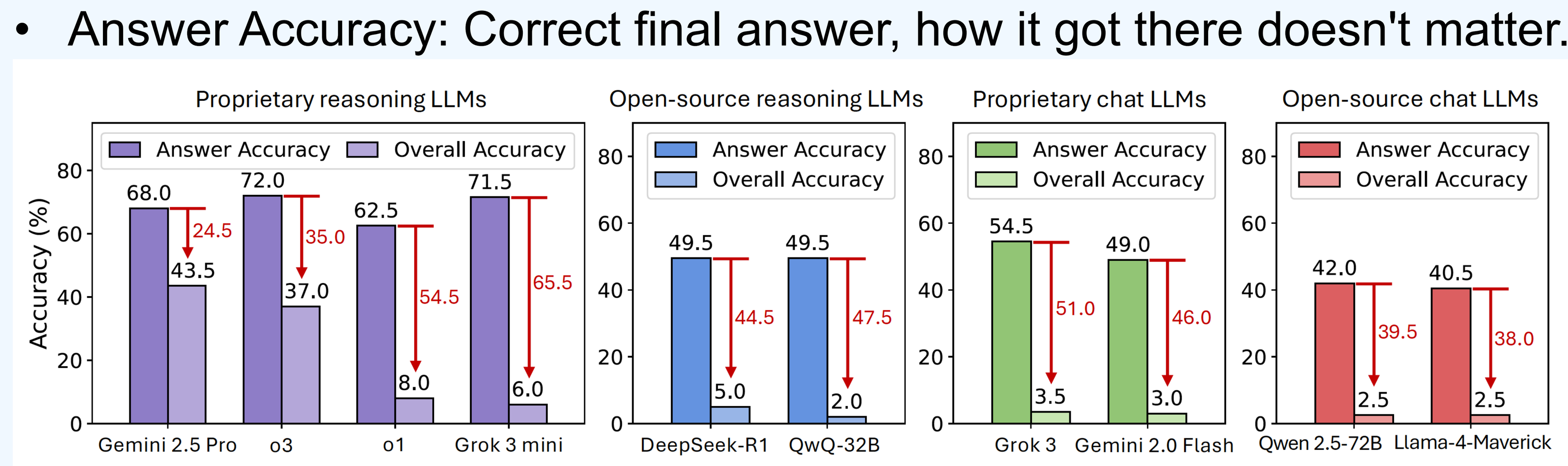


Judge baseline

Key Results

Key Results 1: The "Soundness Gap" is REAL!

- Overall Accuracy: Correct final answer + ALL reasoning steps sound.
- Answer Accuracy: Correct final answer, how it got there doesn't matter.



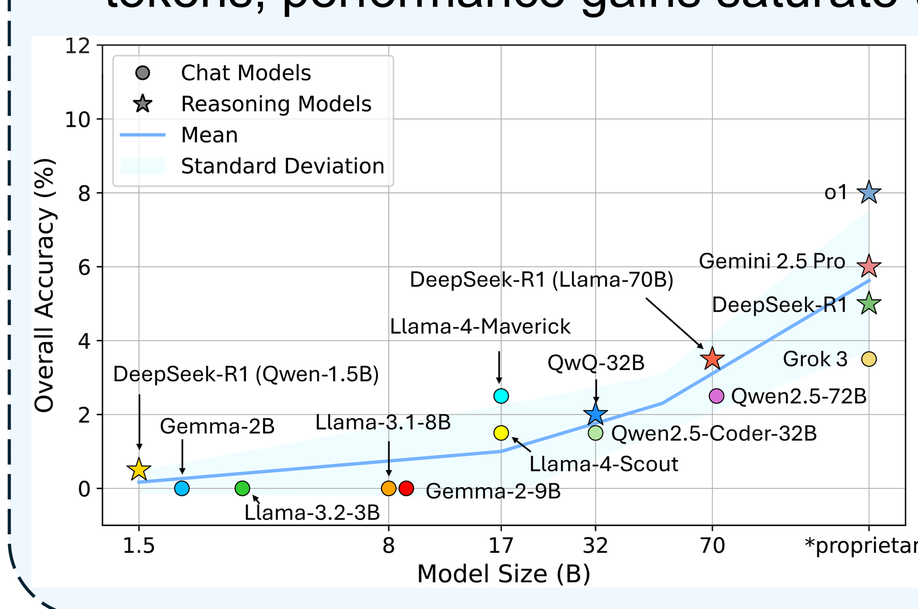
- Overall Acc plummets by **up to 65.5%** compared to Answer Acc.
- This means: LLMs often guess the right answer for complex Olympiad-level inequalities, but their step-by-step reasoning is unsound.

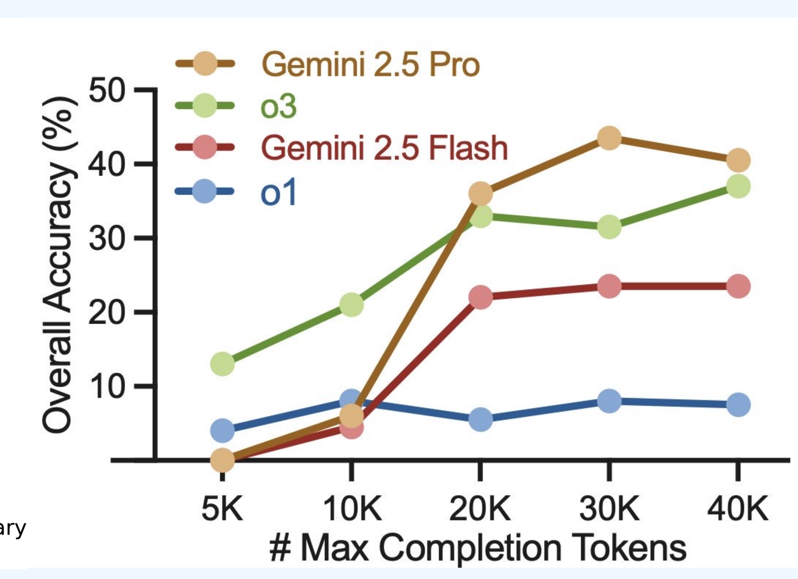
Key Results 2: Model Size Can't solve the Soundness Gap!

- The scaling curve of Overall Accuracy flattens.

Key Results 3: Simply Letting LLMs 'Think' Longer also Doesn't Help.

- While models like Gemini 2.5 Pro and o3 initially improve with more tokens, performance gains saturate (e.g., beyond 20K tokens).



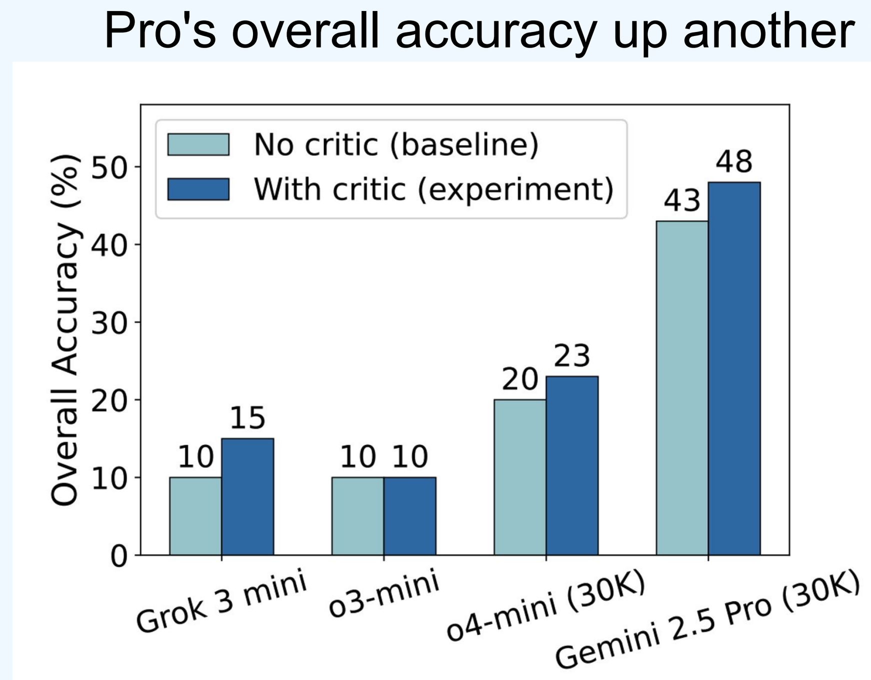


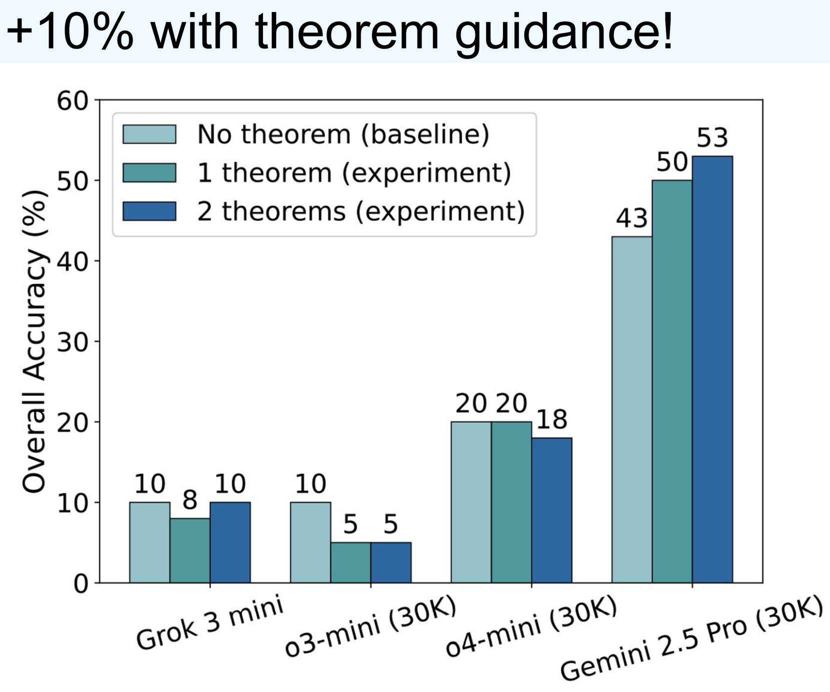
Improvement Strategies

How can LLMs improve their proof rigor on IneqMath?

Two promising paths explored in our work!

1. Self-Improvement (Critic-Guided): Gemini 2.5 Pro's overall accuracy up +5% (43%→48%) via self-critique!
2. Theorem Augmentation (Providing key theorem hints). Gemini 2.5 Pro's overall accuracy up another +10% with theorem guidance!





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