NEURAL INFORMATION PROCESSING SYSTEMS

Stanford

Stanford University Human-Centered Artificial Intelligence

Stanford Artificial Intelligence

IneqMath

Website: https://ineqmath.github.io/

Berkeley UNIVERSITY OF CALIFORNIA

Massachusetts Institute of Technology

Solving Inequality Proofs with Large Language Models

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

[1] Stanford University [2] UC Berkeley [3] MIT | * Co-first authors

Website | Paper | Dataset | Code

Introduction

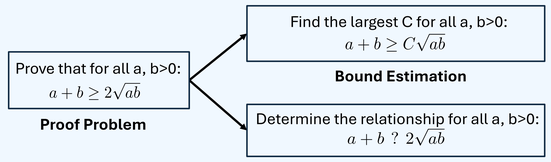
Do large language models truly understand inequality proofs, or do they just make guesses?

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

Goal: Advance LLMs on informal inequality proving.

Task Reformulation

To bridge formal verification with natural language, we reformulate inequality proofs into two **informal** yet **verifiable** subtasks: **Bound Estimation** and **Relation Prediction**.

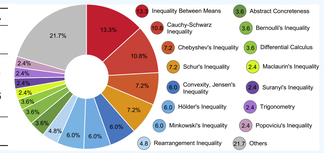


IneqMath Dataset

Based on the reformulation, we create the **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 quality.

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



Dataset Examples

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:

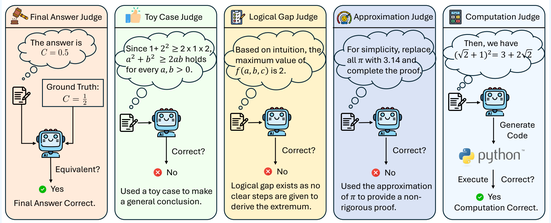
$$3\left(\sum_{cyc} ab\,(1+2\cos(c))\right) \quad (\ ) \quad 2\left(\sum_{cyc}\sqrt{(c^2+ab(1+2\cos(c)))\,(b^2+ac(1+\cos(b)))}\right).$$

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 LLM Judges

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



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**

Dataset Comparison

IneqMath stands out for:

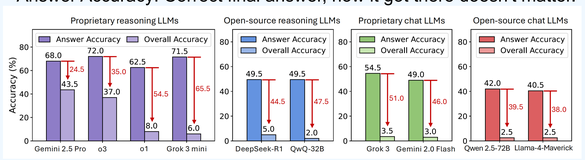
- **Expert-curated** training and testing sets.
- Rich annotations with **step-wise solutions** and 83 grounded **theorems**.
- **Informal** format for inequality proving, evaluated by **LLM judges**.

Datasets	Data Source: Training	Data Source: Test / Dev	Data Annotation: #Theorem	Data Annotation: Solution	Problem and Evaluation: Category	Problem and Evaluation: Format	Problem and Evaluation: Evaluation
INT [64]	Synthesized	Synthesized	35	✓	Proof	Formal	Symbolic DSL
AIPS [63]	Synthesized	✗	8	✓	Proof	Formal	Symbolic DSL
MO-INT [63]	✗	Data compilation	✗	✗	Proof	Formal	Symbolic DSL
MINIF2F [82]	✗	Autoformalization	✗	✗	Proof	Formal	Lean
ProofNet [7]	✗	Autoformalization	✗	✗	Proof	Formal	Lean
FormalMATH [77]	✗	Autoformalization	✗	✗	Proof	Formal	Lean
leanWorkbook [76]	Autoformalization	Autoformalization	✗	✗	Proof	Formal	Lean
Proof or Bluff [49]	✗	Data compilation	✗	✗	Proof	Informal	Human judge
CHAMP [39]	✗	Autoformalization	✗	✗	Open	Informal	Human judge
Putnam Axiom [23]	✗	Data compilation	✗	✗	Open	Informal	Answer checking
LiveMathBench [37]	✗	Data compilation	✗	✗	Open	Informal	Answer checking
IneqMath (Ours)	Expert annotated	Expert annotated	83	✓	MC, Open	Informal	LLM-as-judge

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.



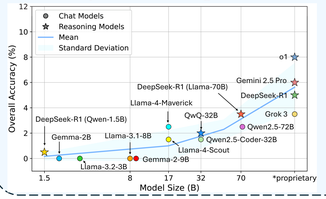
- LLMs often **guess the right answer** for Olympiad-level inequalities, but their **step-by-step reasoning is unsound**.

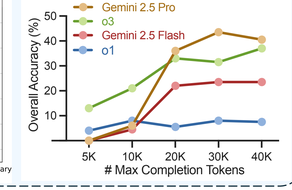
Key Results 2: Bigger Isn't Always Better!

- The increased model size **doesn't enhance overall accuracy**.

Key Results 3: Thinking Longer Can't Solve Soundness Gaps!

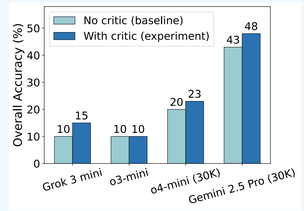
- While models like Gemini 2.5 Pro and o3 initially improve with more tokens, **performance gains saturate** (e.g., beyond 20K tokens).

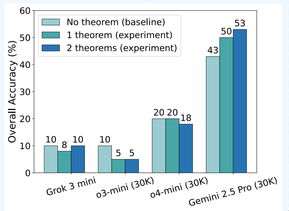




Improvement Strategies

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!

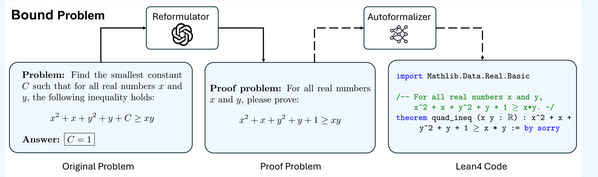


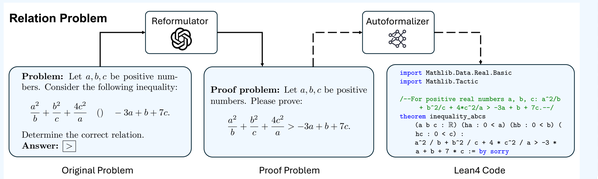


Formal IneqMath Evaluation

- **IneqMath** test problems are reformulated and formalized into **Lean4** code.
- **SOTA** provers achieve only a 14.0% pass rate, showing inequality problems **remain highly challenging**.

Model name	Pass rate (Pass@32)
DeepSeek-Prover-V2-7B [55]	6.0%
Kimina-Prover-Distill-8B [4]	12.0%
Goedel-Prover-V2-32B [36]	13.0%
Goedel-Prover-SFT [35]	14.0%





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