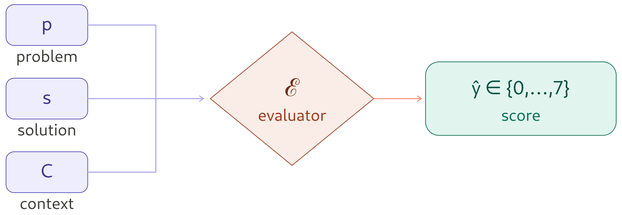
Reliable Fine-Grained Evaluation of Natural Language Math Proofs

Wenjie Ma [1] Andrei Cojocaru [1] Neel Kolhe [1] Haihan Zhang Vincent Zhuang [2] Matei Zaharia [1] Sewon Min [1]

[1]UC Berkeley [2]Google DeepMind

Motivation & the Gap

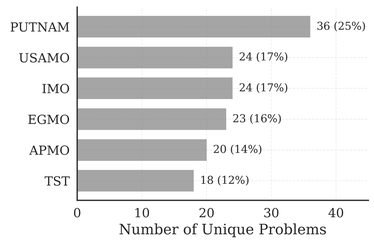
- **The challenge:** LLMs are improving at competition math, but evaluating proofs remains unsolved.
- **Why?** (1) Proof problems may not have a single verifiable final answer. (2) Proof validity needs to check the process.
- **Goal:** Build an **automated fine-grained proof evaluator**.



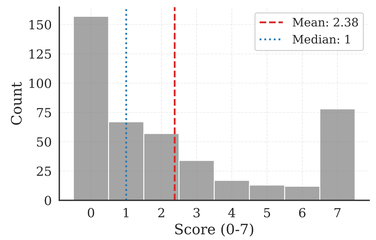
ProofBench: 145 Problems, 435 LLM Solutions

Annotation protocol.

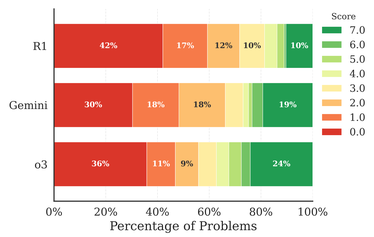
1. **Marking schemes:** LLM-generated, then expert-verified. *Why? Without a clear scheme, partial-credit assignment is highly subjective and inconsistent across solutions.*
2. **Grading:** experts assign a score from 0–7 with the marking scheme.



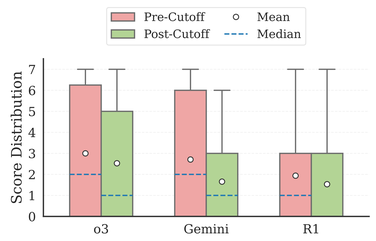
(a) Problem distribution



(b) Overall score distribution



(c) Scores by model

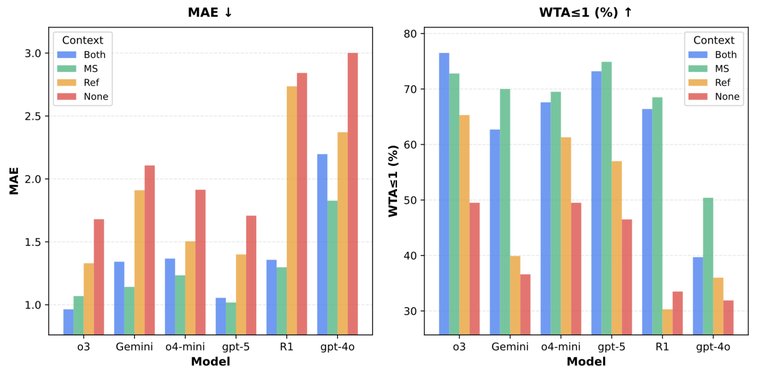


(d) Contamination analysis

SoTA models score $6+$ on $< 30\%$ of solutions

Systematic Study of Evaluator Design

We explored four axes: **backbone**, **context**, **instructions**, and **ensembling**.



- **Backbone strength dominates** — o3 leads across most metrics
- Marking schemes boost all models
- Reference solutions help **selectively**
- Optimal instructions vary by backbone models
- **Best config – ProofGrader:** o3 + ref. sol. + marking scheme + ensemble → MAE 0.926

Example Problem and Marking Scheme

Problem. (USAMO 2025 P2) Let $n > k \geq 1$ be integers. Let $P(x) \in \mathbb{R}[x]$ be a polynomial of degree n with no repeated roots and $P(0) \neq 0$. Suppose that for any real numbers $a_0, \ldots, a_k$ such that $a_k x^k + \cdots + a_1 x + a_0$ divides $P(x)$, the product $a_0 a_1 \ldots a_k$ is zero. *Prove* that $P(x)$ has a nonreal root.

Reference Solution. By considering any $k+1$ roots of P, WLOG assume $n = k+1$. Suppose $P(x) = (x+r_1)\ldots(x+r_n)$ has $P(0) \neq 0$. Then each polynomial $P_i(x) = P(x)/(x+r_i)$ of degree $n-1$ has ≥ 1 zero coefficient. The leading and constant coefficients of each P_i are nonzero, leaving $n-2$ other coefficients. By pigeonhole, P_1 and P_2 share a zero coefficient position, say x^k for some $1 \leq k < n-1$.

Claim. If P_1 and P_2 both have x^k coefficient zero, then $Q(x) = (x+r_3)\ldots(x+r_n)$ has consecutive zero coefficients $b_k = b_{k-1} = 0$.

Proof. By Vieta, let $Q(x) = x^{n-2} + b_{n-3}x^{n-3} + \cdots + b_0$. The x^k coefficient of P_1, P_2 being zero means $r_1 b_k + b_{k-1} = r_2 b_k + b_{k-1} = 0$, hence $b_k = b_{k-1} = 0$ (using r_i nonzero, distinct). □

Lemma. If $F(x) \in \mathbb{R}[x]$ has two consecutive zero coefficients, it cannot have all distinct real roots.

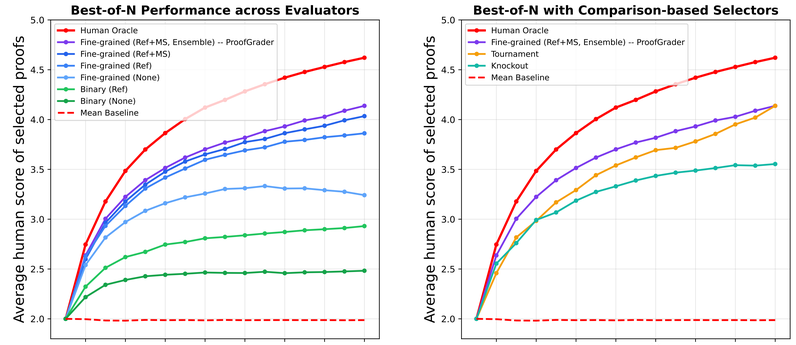
Proof 1 (Rolle). Say x^t, x^{t+1} coefficients are zero. If all roots are real and distinct, Rolle's theorem implies every derivative has this property. But $F^{(t)}(x)$ has a double root at 0, contradiction. □

Proof 2 (Descartes). Real roots are bounded by sign changes in $F(x)$ plus sign changes in $F(-x)$. For consecutive nonzero coefficients $\star x^i, \star x^j$ $(i > j)$: if $i - j = 1$, the sign change counts once; if $i - j \geq 2$, it may count twice but there's ≥ 1 zero between them.
With b nonzero coefficients and z runs of zeros, real roots $\leq b - 1 + z \leq \deg F$. Two consecutive zeros make this strict. □

Marking Scheme (max 7 pts).
Checkpoints (additive):
(1) [1pt] Problem reduction to $n = k+1$ and setup. Alt: complete proof for $k = 1$.
(2) [2pts] Pigeonhole: n polynomials, $n-2$ internal coefficient positions; two share a zero position.
(3) [2pts] Deduce consecutive zeros: from $[x^m]P_1 = [x^m]P_2 = 0$ and $P_i = (x + r_i)Q$, show $b_m = b_{m-1} = 0$.
(4) [2pts] Prove lemma (2pts for complete Rolle or Descartes proof; 1pt for partial).
Deductions: Cap 6/7 if no reduction justification; cap 5/7 if lemma flawed; cap 3/7 if stops after PHP; -1pt for minor gaps (e.g., not using $r_1 \neq r_2$).
Zero credit: Unjustified WLOG; merely stating theorems; specific examples only; noting $P(0) \neq 0 \Rightarrow$ nonzero roots.

Downstream Utility: Best-of-N Selection

We generated 16 candidates per problem on 29 problems from 2025 competitions and compared evaluators by how well they select the strongest proof.



(a) Evaluators (b) Comparison-based

- **ProofGrader: 4.14/7** at $N = 16$, closing **78%** of the gap
- Fine-grained scoring essential: binary judges can't distinguish adequate (5/7) vs. perfect (7/7) proofs
- Outperforms expensive pairwise methods with $\mathcal{O}(N)$ efficiency

Why Fine-Grained Scores?

Case study: IMO 2025 Q4 (expert score: 6/7)

Fine-grained (6/7): verifies sufficiency, proves terms are multiples of 6, correctly analyzes dynamics — only a minor arithmetic slip.

Binary (0/1): one incorrect inequality, one false auxiliary claim — nearly correct proof gets discarded entirely.

Takeaway: Fine-grained scoring preserves relative quality and gives a better signal for reward models. Binary grading collapses partially correct and fully correct proofs into the same label.

Impact: Adopted by External Teams

- **QED-Nano** (CMU, Hugging Face, ETH Zürich, Numina): uses ProofGrader's rubric grading as RL reward signal to train a 4B-parameter Olympiad-level proof model — *no extensive human graders needed.*
- **NVIDIA**: scales generative verifiers to millions of tokens on ProofBench, studying verifier scaling for natural-language proof selection.