From Outcome Evaluation to Process Evaluation: Continuous Capability Scoring for Mathematical Reasoning Benchmarks

Patrick Tenorio

QED Bench

Abstract. Mathematical reasoning benchmarks are still dominated by outcome evaluation: exact-match scoring on contest problems and coarse end-product judgment on proof tasks. These conventions are scalable and reproducible, but they compress each solution into a binary verdict or a small set of rubric checkpoints, obscuring how far a model actually progresses toward a rigorous argument, where it first becomes unsound, and what sort of solution process it can reliably sustain. This problem is especially acute for public contest benchmarks such as AIME, HMMT, and PUMaC. AoPS publicly documents the AIME format and archives AIME problems and solutions, contamination-aware work reports strong signs of contamination in AIME 2024, frontier systems now report perfect or near-perfect AIME 2025 scores, and some public benchmark settings evaluate models in tool-augmented environments with web access [2, 15, 16, 17, 18, 19, 21]. Epoch AI sharpens the distinction by separating contest math from frontier math: in a 2025 review, Grok 4 reaches 88% on a composite of AIME, HMMT, and OTIS-style contests, yet only 12-14% on FrontierMath Tiers 1-3 and 2% on Tier 4 in the same report [31, 33]. Even by March 2026, the best published FrontierMath result remained 50% on Tiers 1-3 and 38% on Tier 4 [32]. At the same time, proof-based and process-based studies continue to show that rigorous mathematical reasoning is far from solved [1, 20]. We therefore propose a process-first rescoring framework that assigns each model a continuous capability score on every problem. Using AIME, HMMT, and PUMaC as both outcome-level and process-level benchmarks, we compare binary exact-match assessment against per-problem capability scoring derived from valid-prefix depth, essential-commitment coverage, unrecoverable-error localization, and self-correction. We then extend the same comparison to curated USAMO and IMO problems, with a longer-term path toward graduate and PhD-level mathematical arguments. Our central claim is that current models remain far from perfect on contest mathematics, that continuous capability scoring provides a more faithful measure of progress toward rigorous argumentation than either binary pass-fail scoring or one-size-fits-all rubrics, and that this evaluation paradigm can guide the field toward more complex mathematical prompts with well-defined solution processes that test sustained derivation rather than end-state answer matching alone.

> Core claim. Exact-match formats should remain as a comparability layer, but each problem should also yield a continuous capability score that measures how far the model progressed toward a rigorous argument. Exhaustive rubrics should be reserved for calibration and auditing, not used as the only scalable evaluation mechanism.

1. Introduction

Mathematical benchmarking for large language models is still dominated by outcome evaluation, most visibly through exact-match contest formats. AIME, for example, is a 15-question, 3-hour examination whose answers are integers from 000 to 999, and many rounds from HMMT and PUMaC inherit the same fixed-answer style [15]. That design is scalable and historically useful, but it compresses a rich reasoning process into a one-bit judgment. A correct final answer does not tell us whether the model reasoned soundly, guessed, repaired a broken derivation late, or followed a memorized template. If the goal is to measure mathematical capability rather than endpoint success alone, outcome evaluation on exact-match contests must be reinterpreted rather than taken at face value [1, 2, 3].

As evaluation has moved from exact-match contests toward olympiad proofs, the gap has become much harder to ignore. Proof or Bluff? evaluated state-of-the-art systems on the 2025 USAMO using expert human graders and found that only one frontier model achieved a non-trivial score, while the others remained below

5% on average [1]. MathArena further argues that popular recurring benchmarks such as AIME 2024 are susceptible to contamination and, by design, do not evaluate proof-writing capability [2]. Epoch AI reaches a compatible conclusion from another direction. Its recent benchmark work explicitly separates medium-hard contest math from frontier math: AIME, HMMT, and OTIS-style contests sit in a much lower difficulty band than FrontierMath's advanced undergraduate through research-level problems [31, 33, 34]. In a 2025 analysis, Epoch reports an 88% composite score for Grok 4 on those contest benchmarks, but only 12-14% on FrontierMath Tiers 1-3 and a single solve out of 48 on Tier 4 [31]. Their March 2026 benchmarking update still places the best FrontierMath result at 50% on Tiers 1-3 and 38% on Tier 4 [32]. Together, these results suggest that current claims about mathematical reasoning are often inflated by scoring conventions and problem selection rather than supported by robust argument construction.

The process-supervision literature offers a constructive basis for moving beyond binary assessment. Earlier verifier work showed that scoring candidate solutions can materially improve mathematical performance [4]. Subsequent comparisons between process-based and outcome-based feedback found that outcome supervision can reduce final-answer error, but process supervision is necessary when the goal is correct reasoning steps rather than merely correct endpoints [5]. Let's Verify Step by Step extended this line by showing that process supervision significantly outperforms outcome supervision on MATH and by releasing PRM800K, a large step-level feedback dataset for training process reward models, or PRMs [6].

This paper presents a benchmark-wide rescoring framework that moves from outcome evaluation to process evaluation. It assigns each model a continuous capability score on every problem, measuring how far the model progresses toward a rigorous argument before becoming unrecoverably unsound. AIME, HMMT, and PUMaC are evaluated under both paradigms so that exact-match assessment and continuous capability scoring can be compared on the same tasks. Curated USAMO and IMO problems then test whether the same evaluator transfers to proof-heavy competition mathematics, with a longer-term extension toward graduate and PhD-level mathematical arguments.

2. Related Work

2.1 Exact-Match Formats and Their Limits

Outcome-level evaluation became dominant for good reasons. It is reproducible, low-cost, and easy to scale to hundreds of problems and many model variants. MATH, AIME-style benchmarks, and many recent leaderboard settings rely on this property. Yet the same simplicity is the main limitation. ReasonEval argues that most mathematical evaluations focus solely on the final result and therefore neglect the validity and redundancy of intermediate steps, masking flaws that are crucial for understanding whether a model truly reasoned its way to the answer [3]. Its empirical conclusion is especially important for benchmark design: higher final-answer accuracy does not necessarily imply higher reasoning quality on challenging problems.

MathArena adds a second objection. Because many well-known competition sets circulate widely online, older answer-only benchmarks are now entangled with contamination risk [2]. This concern is especially acute for AIME. AoPS maintains a long-running public archive of AIME exams and solutions [16], while its AIME overview page documents the exam format and links directly to those resources [15]. MathArena reports strong signs of contamination in AIME 2024 [2]. Meanwhile, public scores are approaching saturation: OpenAI reports 100.0% on AIME 2025 for GPT-5.2 Thinking and GPT-5.2 Pro [17], and 99.5% pass@1 with 100% consensus@8 for o4-mini on AIME 2025 when given a Python interpreter [18]. Some public AIME benchmark settings also evaluate models in tool-augmented agentic loops with shell access and web browsing [19], and frontier model APIs now expose native web search tools [21]. Epoch AI's analyses reinforce why these numbers require caution. It classifies AIME, HMMT, and OTIS as contest math in roughly the 2-6 difficulty band, reports >95% accuracy for public models on competition problems up to difficulty rating 5, yet finds that no LLM has solved even a single problem from the highest tier of USAMO/IMO difficulty [31, 34]. Even on FrontierMath, where contamination is less plausible because the problems are original and private, the best published results remain far from saturation [32, 33]. Yet these numbers should not be mistaken for solved mathematical reasoning. Proof or Bluff? shows that full olympiad proof generation remains weak [1], and OlymMATH warns that similar exact-match gains can reflect

shortcutting and illusory improvement in accuracy rather than rigorous deliberation [20]. This is exactly why a process-driven evaluation is more attractive than a purely outcome-driven one: it asks how much valid mathematics the model actually produced, not just whether it managed to land on the correct integer.

2.2 Process Supervision and Continuous Capability Scoring

Verifier-based work established that intermediate scoring can improve mathematical reasoning. Training Verifiers to Solve Math Word Problems showed that learned verifiers help rank generated solutions and can outperform direct answer generation under scale [4]. Uesato et al. then made the process-versus-outcome distinction explicit: outcome-based supervision produced comparable final-answer error with less label supervision, but process-based supervision was required for correct reasoning steps, reducing reasoning error among final-answer-correct solutions from 14.0% to 3.4% [5].

Let's Verify Step by Step is the canonical reference for process supervision in mathematics. It reported that process supervision significantly outperforms outcome supervision on MATH and released PRM800K with 800,000 step-level correctness labels [6]. This established PRMs as practical tools for selecting and training better reasoning traces, not only for diagnosing them after the fact.

Recent work shifts from training to evaluation. ProcessBench formalizes the task of locating the earliest erroneous step and provides 3,400 human-annotated cases, primarily from competition and olympiad mathematics [7]. The Lessons of Developing Process Reward Models in Mathematical Reasoning warns that PRM pipelines are fragile: Monte Carlo annotation can underperform LLM-as-judge and human annotation, and Best-of-N evaluation can drift back toward outcome-level behavior by rewarding correct answers produced from flawed processes [8]. Beyond the First Error goes further by arguing that first-error-only scoring is too crude for reflective traces, because correct reasoning can reappear after earlier mistakes; it introduces Error Propagation and Error Cessation to distinguish unrecoverable failure from local detours [9].

New benchmark proposals increasingly encode process structure directly. ReasoningMath-Plus introduces minimal reasoning skeletons and reports that answer-only metrics can overestimate reasoning robustness relative to process-aware scoring [10]. This is a particularly important precedent for the present proposal because it suggests that sparse structural annotations can support fine-grained evaluation without requiring exhaustive proof rubrics for every valid solution path.

2.3 Rubric-based evaluation, its rise, and its scaling limits for math

Rubric-based evaluation is becoming a default response to more open-ended and workflow-heavy AI tasks. OpenAI's evaluation guidance recommends detailed scorecards and rubrics for LLM-based grading once exact match and other code-based checks stop capturing the target behavior [24]. OpenAI's recent benchmark design reflects this shift: HealthBench grades 5,000 realistic health conversations with custom physician-created rubrics containing 48,562 rubric criteria [25], frontierscience uses rubric-based grading for open-ended research tasks because short-answer verification trades off against expressivity [26], and the deep research system card describes training and evaluation pipelines that combine ground-truth answers for objective tasks with rubrics for more open-ended ones [27]. Anthropic makes the same move in public guidance and system reports, describing rubric-based scoring as a scalable way to evaluate open-ended agent behavior and using rubric-based evaluation for complex finance workflows involving spreadsheets, slide decks, and documents [28, 29]. Autorubric summarizes this emerging norm directly, arguing that rubric-based LLM evaluation has become standard practice for assessing text generation at scale [30].

This trend is understandable. As tasks become longer, multi-turn, tool-using, and open-ended, exact-match grading becomes too brittle, while rubrics preserve nuance, multidimensionality, and partial credit [24, 25, 26, 28]. But mathematics exposes the limits of rubric scaling more sharply than many application domains. Proof-based evaluation makes the inadequacy of final-answer scoring even clearer. Proof or Bluff? shows that full olympiad proofs remain far harder than short-answer tasks for current models [1]. Reliable Fine-Grained Evaluation of Natural Language Math Proofs presents ProofBench and ProofGrader, and explicitly identifies consistent, scalable expert grading as a central challenge [11]. The paper shows that marking schemes help, but also that graders must treat them as flexible references rather than rigid checklists because valid solution paths vary substantially across proofs.

RefGrader reaches a similar conclusion from a different direction [12]. It finds that models are often able to flag incorrect proofs, including subtle ones, but exhibit calibration gaps when assigning partial credit, so the system derives problem-specific rubrics from reference solutions within a multi-step grading workflow. The design lesson is not that rubrics are misguided. It is that high-quality rubrics for advanced mathematics are difficult to scale. Constructing them requires substantial domain expertise, repeated iteration, and careful anticipation of alternative lemmas, reductions, and case analyses. Because many valid proofs depend on a solver's background, taste, and available toolkit, trying to pre-encode every acceptable path quickly becomes time-consuming, expensive, and incomplete. This is precisely where we suggest a process-evaluation approach: instead of asking annotators to enumerate the full space of valid solutions up front, we score the merits of the particular argument the model actually produced, track how long it remains sound, and measure where it first becomes unrecoverable [7, 9, 10].

3. Why Benchmarks Should Be Rescored

We propose rescoring current mathematical benchmarks rather than abandoning them. Rescoring preserves continuity with established leaderboards while changing what each problem contributes. Instead of assigning every response only a binary pass/fail label or a coarse rubric total, the benchmark should assign a continuous capability score per problem that reflects how much mathematically valid work the model actually completed. Final-answer metrics remain reportable, but they should no longer serve as the primary proxy for reasoning quality.

This shift is especially attractive for competition mathematics because the existing benchmark ecosystem already spans a spectrum from exact-match formats to olympiad proofs. AIME is the canonical exact-match substrate, though public archives, contamination risk, score saturation, and growing tool augmentation make it increasingly unreliable as a standalone proxy for rigorous reasoning [2, 15, 16, 17, 18, 19, 21]. Epoch AI's contest-math analysis reaches a similar conclusion in graded form: public models already exceed 95% on many lower-difficulty contest problems, yet the highest USAMO/IMO difficulty tier remains unsolved by any LLM [34]. HMMT mixes short-answer rounds with a February team round that is proof based, awards partial marks, and includes problems comparable to the hardest USAMO problems [13]. PUMaC combines short-answer individual and team rounds with a Power Round that is explicitly designed to expose students to proof-based mathematics, requires written solutions with justification, and gives teams approximately a week to work [14]. USAMO and the IMO provide the natural proof-heavy endpoint [22, 23], and a longer-term extension can target graduate and research-style mathematical arguments. The missing piece is therefore not a source of problems, but a scoring paradigm that can interpret model capability continuously across this full range.

Rescoring also addresses a practical bottleneck created by the broader shift toward rubric-heavy evaluation. Exhaustive proof rubrics are expensive to author and maintain because difficult problems admit multiple valid strategies, lemmas, and case structures. The right solution path often depends on the solver's technical background, taste, and available mathematical tools, so a single canonical scheme rarely captures the full solution space without repeated expert revision. A capability-scoring framework can instead ask a more portable question: what merits does this particular argument establish, how long does it remain sound, where does it become unrecoverable, and does it later recover? Those questions scale more naturally across heterogeneous solution styles than one-size-fits-all line-by-line rubrics do.

Table 1. Proposed benchmark family, from exact-match formats to olympiad proofs.

Source	**Format**	Outcome-level role	Capability-scoring role	Why included
AIME	15 short-answer integer problems	Exact-answer accuracy, pass@k, self-consistency	Trace scoring on the same exact-match attempts	Preserves continuity with the dominant benchmark style

Source	Format	Outcome-level role	Capability-scoring role	Why included
HMMT individual rounds	Short-answer contest problems	Outcome comparison on broader competition math	Measures latent capability before proof-heavy rounds	Bridges AIME-style tasks and harder contest math
HMMT February team round	Proof-based team round with partial credit	Coarse end-product comparison or normalized score	Step-level soundness, coverage, and fatal-error localization	Natural mixed-format benchmark with olympiad-like proofs
PUMaC individual/team	Short-answer rounds	Outcome comparison on additional competition source	Capability scoring on exact-match tasks	Expands domain coverage and competition style
PUMaC Power Round	Written proof-based problems over about a week	Coarse end-product comparison	Tests long-form reasoning and justification	Captures sustained, collaborative proof-style reasoning
Curated USAMO/IMO set	Six-problem proof exams	Normalized proof-grade endpoint	Highest-rigor capability evaluation	Tests whether capability scores transfer to elite proof generation

4. Benchmark Schema

4.1 Dual scoring on the same problems

The proposed benchmark evaluates the same model traces under two paradigms. For outcome-level evaluation, short-answer tasks receive exact-match scoring and proof tasks receive a normalized end-product score based on coarse human grading or trusted evaluation subsets. For process-level evaluation, the full trace is segmented into reasoning steps and scored for how much of the underlying argument is sound. The key comparison is therefore not between different datasets, but between binary endpoint assessment and continuous capability scoring on the same problems.

Using AIME, HMMT, and PUMaC in both ways is the central methodological move. AIME does not cease to be an exact-answer benchmark. Instead, it becomes a dual benchmark whose same responses can be judged by exact-answer success and by a continuous capability score. HMMT and PUMaC then show how the comparison behaves as problems become less computationally templated and more proof-like. Curated USAMO and IMO problems extend the same logic into a regime where end-product-only judgment is weakest and sustained reasoning is indispensable. Once this interpretation is calibrated on contest mathematics, the same evaluation philosophy can be extended to graduate and research-style mathematical arguments.

4.2 Minimal reasoning skeletons

Each problem is annotated with a minimal reasoning skeleton rather than a full rubric. A reasoning skeleton is a small set of essential commitments needed by any sound solution, up to logically stronger equivalents. Examples include the identification of a correct invariant, the construction of an injective map, a decisive parity reduction, a necessary extremal argument, or a case split that isolates the remaining search space. Annotators are instructed to define the smallest set of commitments that capture what must be established without prescribing the full shape of a canonical proof. This lets the benchmark evaluate the merits of each individual argument rather than forcing every valid solution into one official template.

This choice is motivated by both scalability and fairness. It reduces annotation cost relative to full proof rubrics, and it respects the fact that complex competition problems often admit multiple elegant but structurally different solutions. A model should not be penalized for departing from the official solution so long as it establishes the required mathematical commitments soundly. That flexibility is especially important if the long-term goal is to score proofs and research arguments whose valid presentations may differ substantially across solvers and subfields.

Table 2. Minimal annotation objects for each problem.

Object	Definition	Used for
Step segmentation	Coherent reasoning unit such as a derived claim, case split, reduction, or lemma	Local validity and progress judgment
Essential commitment	Minimal claim that any sound solution must establish, or replace with a stronger equivalent	Coverage-based scoring across different solution paths
Recovery marker	Point at which a model returns to a sound trajectory after a local error	Self-correction credit and reflective traces
Unrecoverable error	Step after which the current argument cannot succeed without changing a core idea	Valid-prefix depth and failure localization
Outcome tag	Final answer or end-product status	Direct comparison between process and outcome metrics

5. Annotation Protocol

The benchmark requires a principled annotation protocol because continuous capability scoring becomes inconsistent if step boundaries or error definitions are vague. We propose a two-layer annotation pipeline. In the first layer, experts create minimal reasoning skeletons and identify acceptable equivalent commitments. In the second layer, annotators score model traces step by step using a common label set and a decision policy for local mistakes, recoveries, and unrecoverable failures.

Step segmentation should follow mathematical intent rather than raw sentence boundaries. A single step may contain several sentences if they express one atomic claim, and a sentence may be split if it contains multiple logically distinct moves. Annotators should prefer segments that can be judged as sound, unsupported, or erroneous without needing to partially split the claim further. The guiding principle is that each step should correspond to a unit a human grader could reasonably validate or reject.

Table 3. Step-level label schema.

Label	Meaning	Metric impact
Sound-progressing	Mathematically valid and moves the solution forward	Increases valid-prefix depth and commitment coverage
Sound-nonprogressing	Valid but redundant, circular, or not advancing the solution	Counts as sound but may reduce efficiency analyses
Unsupported but repairable	Claim may be salvageable with a missing justification or small patch	Marks a weakness but not immediate fatal failure
Material error	Incorrect claim, invalid inference, or broken computation that affects the argument	Candidate error point requiring recovery analysis

Label	Meaning	Metric impact
Unrecoverable error	Failure after which the current trajectory cannot succeed without replacing a core idea	Terminates the valid prefix for primary scoring
Recovery	Subsequent step that returns the model to a mathematically sound path	Adds self-correction credit when justified

To keep the benchmark scalable, full expert labeling is not required for every trace. Instead, a calibrated subset should be double-annotated by experts, with disagreements adjudicated to create a gold slice for PRM or critic-model validation. Automatic process evaluators can then score the larger benchmark, but only after they are benchmarked against the adjudicated subset. This design treats rubrics and expert proof grading as calibration resources rather than as the universal scoring interface for all samples, which is essential if per-problem capability scores are to scale beyond a small curated set.

Proof traces require one additional rule: equivalent strategic routes are acceptable. If a solution reaches the required commitments through a different but sound decomposition, annotators should map it to the nearest equivalent commitment rather than scoring it as off-rubric. This is exactly where minimal reasoning skeletons improve over rigid marking schemes, because they reward the merits of the argument that was actually produced.

6. Continuous Capability Scoring

Let O denote the normalized outcome score. On short-answer tasks, O is exact final-answer correctness and takes values in {0,1}. On proof tasks, O is a normalized end-product score derived from a coarse human grade or validated proof evaluator. Let V denote valid-prefix ratio, defined as the proportion of steps before the first unrecoverable error. Let C denote essential-commitment coverage, the fraction of problem commitments established soundly. Let R denote recovery credit, a normalized measure of whether the model returns to a sound path after local mistakes.

We define the per-problem capability score K as a convex combination of valid-prefix depth, commitment coverage, and recovery credit. The precise weights should be tuned on the expert-adjudicated development split to maximize agreement with human judgments. The benchmark then reports a capability-outcome divergence statistic D = O - K, along with an unsupported-correctness rate that measures how often exact-answer-correct solutions still have low capability scores. These statistics directly quantify where binary pass/fail benchmarks overstate reasoning quality.

V = (steps before first unrecoverable error) / (total steps)

C = (essential commitments established soundly) / (total essential commitments)

K = alpha V + beta C + gamma R, where alpha + beta + gamma = 1

D = O - K

U(tau) = Pr(O = 1 | K < tau)

The benchmark leaderboard should rank models primarily by average capability score K-bar, not by O. Outcome-level metrics remain visible because they preserve continuity with prior work. But the principal scientific question becomes how far a model progresses toward a rigorous argument before failure, not merely whether the final line matches an answer key or a response happens to satisfy a fixed set of rubric checkpoints.

7. Hypotheses and Conjectures

Hypothesis 1: Outcome-capability divergence grows with problem difficulty. Rankings by exact-answer accuracy and continuous capability score should remain closer on AIME than on HMMT February, PUMaC Power Round, and curated USAMO and IMO problems. This follows from ReasonEval's finding that final-answer improvements do not necessarily track reasoning quality, from proof-centric evaluations showing that models degrade sharply when full justification becomes necessary, and from Epoch AI's distinction between strong contest-math performance and much weaker frontier-math performance [1, 3, 31, 32, 34].

Hypothesis 2: Unsupported correctness is common on exact-match contest formats. A non-trivial portion of AIME, HMMT, and PUMaC traces that end with the correct answer should still receive low capability scores because the derivation contains missing justifications, invalid leaps, or an unrepaired fatal step. This is the benchmark-level analogue of the reasoning-error findings reported by Uesato et al. and the process-outcome misalignment documented in PRM evaluation work [5, 8].

Hypothesis 3: Wrong answers often hide substantial latent capability. Many unsuccessful traces should still contain long valid prefixes and high commitment coverage before a decisive error. Exact-match scoring discards that signal, while capability scoring preserves it. ProcessBench and Beyond the First Error both motivate this claim by focusing on earliest error localization and the possibility of meaningful self-correction [7, 9].

Hypothesis 4: Capability scores transfer to proof generation better than outcome scores. Models with higher capability scores on AIME, HMMT, and PUMaC should predict stronger performance on curated USAMO and IMO proofs more faithfully than models with higher exact-answer accuracy alone. If this holds, it would show that continuous per-problem capability scoring is not just more descriptive, but also more predictive of high-rigor reasoning and a better foundation for future evaluation of graduate and research-style arguments.

Hypothesis 5: Minimal reasoning skeletons scale better than full rubric construction. Sparse structural annotations, combined with a calibrated PRM or critic model, should reach strong agreement with human evaluation at much lower annotation cost than bespoke proof rubrics. This conjecture is justified by ReasoningMath-Plus on the one hand and by ProofGrader and RefGrader on the other: the former shows the promise of minimal reasoning structure, while the latter show how costly rich rubric-based grading becomes at scale [10, 11, 12].

8. Experimental Plan

The benchmark should be built in four stages. Stage 1 collects problems and official or widely used archival solutions from AIME, HMMT, and PUMaC, then curates a temporally controlled USAMO and IMO proof subset [13, 14, 15, 16, 22, 23]. Stage 2 creates minimal reasoning skeletons and expert annotations for a calibrated development split. Stage 3 scores traces from a range of frontier and open models under both outcome-level and capability-scoring paradigms. For short-answer tasks, each trace will receive exact-answer correctness and process labels. For proof tasks, each trace will receive a coarse end-product grade and the same process labels. Stage 4 pilots the same scoring philosophy on a small set of graduate and research-style mathematical arguments to test how far the benchmark can generalize once calibrated on contest mathematics.

The first empirical analysis should compare model rankings under O and K across benchmark slices. The second should quantify unsupported correctness and latent capability through D and U(tau). The third should test predictive transfer: does a model's capability score on AIME, HMMT, and PUMaC predict proof performance on curated USAMO and IMO tasks better than its exact-answer accuracy does? A fourth analysis should examine whether the same evaluator remains informative on pilot graduate and research-style problems. Together, these analyses would demonstrate not only that binary and continuous scoring differ, but also that continuous capability scoring better captures the kind of reasoning the field ultimately cares about.

9. Limitations and Scope

A capability-scoring benchmark does not eliminate the need for human mathematical judgment. Automatic evaluators can fail, especially on subtle proofs and long reflective traces. Existing PRM work already warns that data synthesis, Best-of-N evaluation, and naive first-error assumptions can distort what a capability score means [7, 8, 9]. For this reason, human calibration remains essential.

The proposal also does not claim that rubric-based evaluation is a mistake. On the contrary, frontier labs increasingly rely on rubrics for open-ended agent, research, and domain-specific evaluations [24, 25, 26, 27, 28, 29, 30], and in high-stakes proof grading they remain valuable [11, 12]. The narrower claim is that exhaustive rubrics should not be the default scalable interface for benchmark-wide rescoring, especially when complex mathematical problems admit many valid solution paths. Their role should be to anchor calibration, adjudicate difficult cases, and validate the automated evaluator.

10. Conclusion

Mathematical reasoning benchmarks still overweight the final line of a solution. The literature on process supervision, process evaluation, contamination-aware benchmarking, rubric-based evaluation, and proof grading converges on a common lesson: correct answers and correct reasoning are not the same object of evaluation. This lesson is especially urgent for AIME-style exact-match benchmarks, where public archives, contamination concerns, native web search, tool-augmented benchmarking, and now perfect or near-perfect public scores make binary outcome metrics increasingly hard to interpret [2, 15, 16, 17, 18, 19, 21]. Epoch AI's recent work makes the same point by distinguishing contest math from frontier math: models can score very highly on AIME, HMMT, and OTIS-style contests while remaining far from saturated on FrontierMath and the hardest USAMO/IMO tiers [31, 32, 33, 34]. We therefore advocate a transition from outcome evaluation to process evaluation: rescore AIME, HMMT, and PUMaC with continuous per-problem capability scores, then extend the same comparison to curated USAMO and IMO proof sets. The immediate goal is to show that current model capabilities on contest mathematics are far from perfect once responses are interpreted continuously rather than as pass/fail events. The longer-term goal is to use the same evaluation philosophy to study graduate and research-style mathematical arguments without relying on a one-size-fits-all rubric. Just as importantly, this paradigm should reveal the need for more complex mathematical prompts with well-defined solution processes, because once we can localize where exact-match contest traces break down, benchmark design no longer has to stop at final answers or coarse checkpoints. It can instead require models to sustain explicit chains of justified commitments over longer arguments, creating a principled bridge from contest math to olympiad proofs and eventually to research-level reasoning. If successful, this framework would provide a more faithful basis for measuring progress in mathematical reasoning and for evaluating the merits of whatever argument a model actually produces.

References

[1] I. Petrov et al. Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad. arXiv:2503.21934, 2025.

[2] M. Balunović et al. MathArena: Evaluating LLMs on Uncontaminated Math Competitions. arXiv:2505.23281, 2025.

[3] S. Xia et al. Evaluating Mathematical Reasoning Beyond Accuracy. arXiv:2404.05692, 2024.

[4] K. Cobbe et al. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168, 2021.

[5] J. Uesato et al. Solving Math Word Problems with Process- and Outcome-Based Feedback. arXiv:2211.14275, 2022.

[6] H. Lightman et al. Let's Verify Step by Step. arXiv:2305.20050, 2023.

[7] C. Zheng et al. ProcessBench: Identifying Process Errors in Mathematical Reasoning. arXiv:2412.06559, 2024.

[8] Z. Zhang et al. The Lessons of Developing Process Reward Models in Mathematical Reasoning. arXiv:2501.07301, 2025.

[9] Z. Yang et al. Beyond the First Error: Process Reward Models for Reflective Mathematical Reasoning. arXiv:2505.14391, 2025.

[10] X. Zheng et al. Unmasking Reasoning Processes: A Process-aware Benchmark for Evaluating Structural Mathematical Reasoning in LLMs. arXiv:2602.00564, 2026.

[11] W. Ma et al. Reliable Fine-Grained Evaluation of Natural Language Math Proofs. arXiv:2510.13888, 2025.

[12] H. Mahdavi et al. RefGrader: Automated Grading of Mathematical Competition Proofs using Agentic Workflows. arXiv:2510.09021, 2025.

[13] Harvard-MIT Mathematics Tournament. Testing Information. Official tournament page.

[14] Princeton University Mathematics Competition. Competition Rules. Official tournament page.

[15] Art of Problem Solving. American Invitational Mathematics Examination. AoPS Wiki.

[16] Art of Problem Solving. AIME Problems and Solutions. AoPS Wiki.

[17] OpenAI. Introducing GPT-5.2. Official release page.

[18] OpenAI. Introducing o3 and o4-mini. Official release page.

[19] Artificial Analysis. AIME 2025 Benchmark Leaderboard. Evaluation page.

[20] H. Sun et al. An Olympiad-Level Math Benchmark for Large Language Models. arXiv:2503.21380, 2025.

[21] OpenAI. Web search guide. Official API documentation.

[22] Mathematical Association of America. MAA Invitational Competitions. Official competition page.

[23] International Mathematical Olympiad. General Regulations. Official regulations document.

[24] OpenAI. Evaluation best practices. Official API documentation.

[25] OpenAI. Introducing HealthBench. Official benchmark page.

[26] OpenAI. Evaluating AI's ability to perform scientific research tasks. Official benchmark page.

[27] OpenAI. Deep research system card. Official system card.

[28] Anthropic. Demystifying evals for AI agents. Official engineering post.

[29] Anthropic. Claude Sonnet 4.6 System Card.

[30] D. Rao and C. Callison-Burch. Autorubric: A Unified Framework for Rubric-Based LLM Evaluation. arXiv:2603.00077, 2026.

[31] G. Burnham. Evaluating Grok 4's math capabilities. Epoch AI report, 2025.

[32] Epoch AI. AI Capabilities benchmarking hub and FrontierMath benchmarking updates. Official benchmarking page, 2026.

[33] T. Besiroglu, E. Glazer, and C. Falkman Olsson. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI. arXiv:2411.04872, 2024.

[34] G. Burnham. LLMs have not yet solved the hardest problems on high school math contests. Epoch AI data insight, 2025.