Toward Unified Evaluation of Mathematical AI: A Survey and Framework

Nishitha A [1]

Abstract

Large language models have achieved remarkable results on mathematical reasoning tasks, yet the field lacks a unified framework for evaluating both natural-language solutions and formal proofs. Existing surveys treat formal verification, LLM-as-judge, process reward models, and human grading as isolated paradigms organized by training pipeline or dataset role, while the 2024–2026 period has produced a fragmented landscape of over thirty new benchmarks spanning competition mathematics, undergraduate coursework, and research-level problems. We present a paradigm-centric taxonomy crossing the formal/informal boundary, organizing evaluation methodology into six paradigms—symbolic checkers, formal proof verifiers, LLM-as-judge, process/outcome reward models, human expert grading, and hybrid generative verifiers—and systematically classify benchmarks along five orthogonal axes: difficulty tier, answer type (final-answer vs. proof), formalization level, contamination resistance, and linguistic breadth. We then provide direct evidence that paradigms diverge systematically: a synthesis of published results and a confirmatory study on the Open Proof Corpus (n=150) yield cross-paradigm Cohen's κ in the 0.04–0.28 range, with correlated over-acceptance between LLM-judge and PRM signals in 48% of disagreements. Our critical gap analysis reveals that no existing verifier paradigm simultaneously satisfies correctness guarantees, scalability, and alignment with human mathematical judgment, motivating a research agenda toward unified evaluation.

[1]Computer Science (AI & ML), PES University, Bengaluru, India. Correspondence to: Nishitha A <nishithaanand2004@gmail.com>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Mathematical reasoning has emerged as a central testbed for large language model (LLM) capability, with performance on olympiad-level problems approaching—and in some domains exceeding—median human expert performance (Luong et al., 2025). Yet despite rapid progress on *solving* mathematical problems, the question of how to *evaluate* those solutions remains surprisingly unsettled.

The core difficulty is that mathematics admits two qualitatively distinct output types. **Final-answer problems**—where the model produces a numerical or symbolic answer—admit relatively straightforward evaluation via symbolic equivalence checking. **Proof problems**—where the model must produce a complete, logically valid argument—require assessing correctness, rigor, mathematical insight, and presentational clarity. Historically these two evaluation problems have been treated separately: the formal theorem-proving community evaluates via proof assistants (Lean, Isabelle, Coq/Rocq); the natural-language reasoning community uses LLM-as-judge, rubric-based human scoring, or process reward models (PRMs). No survey has unified them around the verification paradigm as the primary axis.

This fragmentation is becoming a serious obstacle. As models begin tackling genuine mathematical research problems (Glazer et al., 2024), evaluators need benchmarks and verifiers that generalize across both final-answer and proof settings, across formal and natural-language representations, and across difficulty tiers. The proliferation of benchmarks—our survey identifies over thirty released between 2024 and early 2026—has not been accompanied by corresponding methodological synthesis.

Contributions. We make four primary contributions: (1) a *paradigm-centric taxonomy* of evaluation methodology crossing the formal/informal boundary, organized around the verifier rather than the training pipeline or dataset role; (2) a systematic *benchmark classification* along five orthogonal axes for the 2024–2026 generation of datasets; (3) *direct empirical evidence of cross-paradigm disagreement*—a synthesis of published results plus a confirmatory study on the Open Proof Corpus comparing three paradigms on a shared sample of 150 human-graded proofs; and (4) a *critical gap analysis and research agenda* for next-generation verification.

Scope. We focus on evaluation and verification methodology, not on model architectures for solving mathematics. We do not survey prompting strategies, SFT pipelines, or RL training objectives except insofar as they bear directly on evaluation. *What this paper does not do:* we do not propose new verifiers, retrain existing ones, or evaluate frontier closed-source models; we map the verifier landscape and present a single confirmatory study illustrating that the landscape's divergence is observable empirically.

Notation. We use the following acronyms throughout: **PRM** (process reward model, scores intermediate reasoning steps); **ORM** (outcome reward model, scores final answers); **BoN** (best-of-N sampling/selection); **CoT** (chain-of-thought); **RLVR** (reinforcement learning from verifier rewards); **DPO** (direct preference optimization); **AUC-PR** (area under the precision–recall curve); **GenSelect** (generative tournament selection, a hybrid verifier pattern); **ICC** (intraclass correlation coefficient for partial-credit agreement); **IRT** (item response theory).

2. Background and Adjacent Surveys

Eight recent surveys partially address our scope; none organizes the literature around the verification paradigm as the primary axis. Table 1 summarizes their coverage across four pillars of our framework; two cuts are the most informative contrast. Li et al. (2024) cover formal systems and their metrics but predate the 2025 explosion of natural-language proof evaluation (USAMO/IMO proof grading, Open Proof Corpus, ProofBench, IMO-ProofBench, MathOlympiadBench). Amjad et al. (2026) organize datasets by their role in the training pipeline (pretraining vs. SFT vs. evaluation); our paradigm-centric cut is orthogonal and foregrounds the formal/informal boundary that role-based organization hides. Ahn et al. (2024), Wang et al. (2025), Yan et al. (2025), Wei et al. (2025), and Zheng et al. (2025) treat verifiers and judges only instrumentally—inside an RLVR or PRM-training context—rather than as a primary object of study. Luong et al. (2025) introduce IMO-Bench as a primary research contribution that instantiates one paradigm pair (LLM-as-judge calibrated against human experts), but does not taxonomize the broader paradigm space.

The paradigm-centric unification across the formal/informal evaluation boundary, and the taxonomization of the 2025 benchmark surge, constitute the novel framing of this survey.

3. A Paradigm-Centric Taxonomy of Verification

We organize evaluation methodology into six paradigms, ordered roughly by the correctness-guarantee vs. scalability trade-off (Figure 1).

Table 1. Coverage of adjacent surveys and integrated frameworks. ✓= primary focus; ○ = partial; × = absent.

Work	Formal	NL Proof	PRM	Bench.
Li et al. (2024)	✓	×	×	○
Ahn et al. (2024)	×	×	×	✓
Wang et al. (2025)	○	×	○	✓
Yan et al. (2025)	×	×	×	✓
Wei et al. (2025)	×	×	✓	○
Zheng et al. (2025)	×	×	✓	○
Amjad et al. (2026)	○	○	○	✓
Luong et al. (2025)	×	✓	×	○
Ours	✓	✓	✓	✓

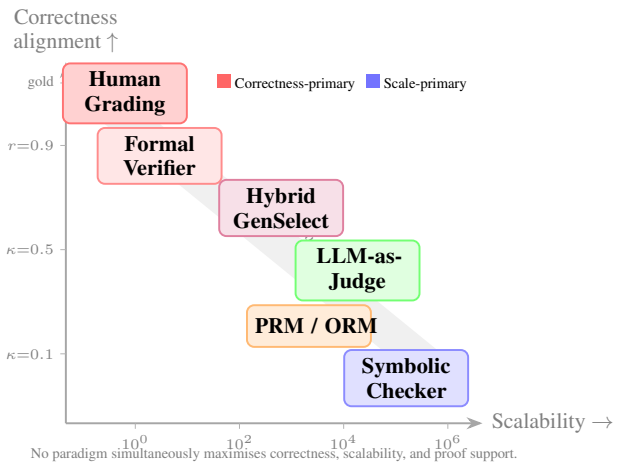

Figure 1. Six verification paradigms on the correctness-alignment vs. scalability trade-off. Horizontal axis: approximate throughput (proofs evaluable per day on commodity hardware, log scale, derived from the cost column of Table 2). Vertical axis: agreement with human expert grading, anchored at published values: $\kappa{\approx}0.13$ for open-source 7B judges (Section 6.2); $\kappa{\approx}0.5$ for fine-tuned judges (Dekoninck et al., 2025); $r{\in}[0.87, 0.96]$ for frontier judges with structured rubrics (Luong et al., 2025; Dekoninck et al., 2025; Yue & Klein, 2025); gold for human grading. No paradigm occupies the top-right corner; the shaded band marks the Pareto frontier.

3.1. Paradigm 1: Symbolic Answer Checkers

Tools such as SymPy-based equivalence checking (Meurer et al., 2017), LightEval (Fourrier et al., 2024), and SimpleRL verify final-answer correctness by symbolic manipulation. They are fast, deterministic, and fully automated. However, they exhibit well-documented brittleness to format and representation differences (e.g., $\frac{a}{b}$ vs. a/b, trigonometric simplifications), and are inapplicable to open-ended proofs.

Verdict. Appropriate for controlled-format competition final-answer benchmarks; unreliable as a sole evaluation mechanism for harder or open-ended problems.

3.2. Paradigm 2: Formal Proof Verifiers

Proof assistants—Lean 4, Isabelle, Coq/Rocq, HOL Light, Metamath—provide *correctness guarantees* by construction. The trade-off is the *autoformalization bottleneck*: natural-language proofs must first be translated to a formal representation, which remains imperfect—Ospanov et al. (2025) document that over half of MiniF2F-v1's formal statements were misaligned with their natural-language counterparts. Recent results illustrate the scalability cost: DeepSeek-Prover-V2-671B reaches 88.9% on MiniF2F-test only at Pass@8192 (Ren et al., 2025); Goedel-Prover-V2 (Lin et al., 2025) solves 86 PutnamBench problems at Pass@184 with a 32B model; and RLMEval (Poiroux et al., 2025) on research-level Lean blueprint theorems caps the best model at 10.3% pass rate.

Verdict. Correctness-guaranteeing but compute-intensive and bottlenecked by autoformalization quality.

3.3. Paradigm 3: LLM-as-Judge for Natural-Language Proofs

Using a frontier LLM to assess natural-language proofs has emerged as a scalable alternative to human expert grading. Yang et al. (2025) demonstrate applicability to ProofBench (APMO/EGMO/IMO/Putnam/USAMO); Dekoninck et al. (2025) show in OPC (>5,000 human-graded LLM proofs) that judge fine-tuning closes much of the gap to expert evaluation; and Luong et al. (2025) report r=0.93–0.96 on IMO-ProofBench with Gemini-2.5-Pro plus detailed guidelines, dropping to 0.87 on a 170-system pool. Yue & Klein (2025) corroborate: rubric structure is load-bearing, with near-random performance once it is removed. JudgeBench (Tan et al., 2024) supplies meta-evaluation infrastructure for judge reliability.

Verdict. Scalable; reliable only with rubrics and fine-tuning; requires meta-evaluation in new domains.

3.4. Paradigm 4: Process and Outcome Reward Models

Process reward models (PRMs) assign step-level correctness signals, enabling both training (RLVR, DPO) and inference-time selection (best-of-N). PRM800K (Lightman et al., 2023), Math-Shepherd (Wang et al., 2024), and Qwen2.5-Math-PRM-7B/72B (Zhang et al., 2025) represent the state of the art. However, PRMs exhibit documented biases in best-of-N selection; Zhang et al. (2025) report systematic failure modes where step-level rewards misalign with final solution correctness. Outcome reward models (ORMs) are coarser but more robust for final-answer tasks.

Verdict. Strong as training signals; biased at inference time; not directly applicable to formal proof tasks.

3.5. Paradigm 5: Human Expert Evaluation

Human expert grading remains the gold standard for mathematical quality, insight, and presentational rigor. GHOSTS (Frieder et al., 2024), the USAMO proof study (Petrov et al., 2025) (only Gemini-2.5-Pro scored $\geq$25%; others <5%), and IMO-GradingBench (Luong et al., 2025) (1,000 human-graded proofs) collectively demonstrate both the difficulty of current benchmarks and human grading's reliability as a reference signal. The Open Proof Corpus (Dekoninck et al., 2025) is the largest open dataset to date (>5,000 graded proofs across >1,000 problems).

Verdict. Currently irreplaceable for ground truth and calibration; expensive and slow.

3.6. Paradigm 6: Hybrid Generative Verifiers

The most recent paradigm (Liu et al., 2025; Liang et al., 2025) combines LLM-as-judge with tournament-style selection (GenSelect) or decoupled reasoning/proving pipelines, providing both scalable proof assessment and test-time compute scaling. This represents the current frontier for natural-language proof evaluation, enabling evaluation at the scale of formal verifiers without the autoformalization requirement.

Verdict. Promising synthesis of scalability and reliability; calibration against formal and human baselines is ongoing.

Table 2 provides a concise comparison of all six paradigms across four key dimensions.

Table 2. Comparison of six verification paradigms. ✓= satisfies; ○ = partial; × = does not satisfy.

Paradigm	Correct.	Scale	Proof	Cost
Symbolic checker	○	✓	×	Low
Formal verifier	✓	○	✓	High
LLM-as-judge	○	✓	✓	Med
PRM/ORM	○	✓	○	Med
Human expert	✓	×	✓	V. High
Hybrid (GenSelect)	○	✓	✓	Med

4. Evaluation Metrics Across Paradigms

Beyond the choice of *verifier paradigm*, evaluation requires selecting appropriate *metrics*. Their incompatibility across paradigms is itself a major source of field fragmentation.

Pass@k and variants. For formal proof tasks, Pass@k measures the probability that at least one of k sampled proof attempts verifies. Pass@k conflates model reliability with inference-time compute: Pass@8192 vs. Pass@32 yields dramatically different impressions of capability (88.9% vs. 82.4% on MiniF2F (Ren et al., 2025)). We recommend that leaderboards always report the sample budget k alongside

the headline number.

Exact match and symbolic equivalence. For final-answer tasks, exact string match is brittle; symbolic equivalence (SymPy, Mathematica) is preferred but has documented failure modes on trigonometric identities, piecewise functions, and non-standard representations. Majority-voting accuracy (self-consistency) is a more robust alternative.

Rubric-based scores. For natural-language proofs, human experts assign partial credit according to marking schemes (typically 0–7 per problem for IMO). LLM-as-judge systems replicate this, with inter-annotator agreement (Cohen's κ, ICC, or Krippendorff's α) against human experts as the key calibration metric.

Step-level accuracy for PRMs. PRMs are evaluated by step-level F1 and AUC-PR for first-error detection. Zhang et al. (2025) show high AUC-PR does not reliably translate to high best-of-N accuracy when reward-hacking trajectories exist—high-precision step rewards can co-exist with globally incorrect conclusions.

Autoformalization alignment. For Paradigm 2, the BEqL/BEq+ indices (Poiroux et al., 2025) quantify how faithfully an autoformalized proof preserves the original argument's content, decoupling autoformalization quality from downstream proof search.

5. Benchmark Taxonomy

We classify 2024–2026 benchmarks along five axes; Table 3 summarizes the most significant entries.

Axis 1: Difficulty tier. Competition (AMC/AIME/Olympiad), undergraduate, and research-level tiers. Research-level benchmarks such as FrontierMath (Glazer et al., 2024) and RLMEval (Poiroux et al., 2025) expose the frontier of model capability.

Axis 2: Answer type. Final-answer benchmarks (GSM8K, MATH, AIME, OmniMATH) vs. proof-requiring benchmarks (MiniF2F, PutnamBench, Open Proof Corpus, ProofBench). The 2025 emergence of natural-language proof benchmarks marks a qualitative shift in evaluation difficulty.

Axis 3: Formalization level. Fully formal (Lean/Isabelle/Coq; MiniF2F, LeanDojo, PutnamBench, FormalMATH), autoformalization-requiring, or natural-language (USAMO, OPC, ProofBench, IMO-ProofBench).

Axis 4: Contamination resistance. Static benchmarks (fixed problem sets, risk of training-set contamination) vs. live/dynamically generated benchmarks (MathArena (Balunović et al., 2025) with live competition problems, FrontierMath with private problems).

Axis 5: Linguistic breadth. English-only vs. multilingual. OlympiadBench (He et al., 2024) addresses cultural and linguistic generalization with bilingual problems.

Saturation analysis. Final-answer competition benchmarks (GSM8K, MATH) are effectively saturated. MiniF2F is approaching saturation at high sample budgets (88.9% at Pass@8192 (Ren et al., 2025)) but not at practical ones (82.4% at Pass@32). Natural-language olympiad proof benchmarks remain far from saturation: only Gemini-2.5-Pro exceeded 25% on USAMO 2025 (Petrov et al., 2025); best non-Gemini models trail IMO-ProofBench by 42.4 percentage points (Luong et al., 2025).

6. Cross-Paradigm Disagreement

A core claim of this paper is that different verification paradigms produce systematically divergent verdicts on the same model outputs. We substantiate this claim in two complementary ways: first, by synthesizing published results across paradigms on overlapping problem sets (Section 6.1); second, by a small confirmatory experiment comparing three paradigms on a shared sample of human-graded proofs (Section 6.2).

6.1. Evidence from Published Results

Formal vs. natural-language proof evaluation. The starkest evidence of cross-paradigm divergence comes from comparing formal-verifier pass rates with informal evaluation scores on matched olympiad problems. Dekoninck et al. (2025) report that on the PutnamBench subset of OPC, the best formal prover (Goedel-Prover-V2 (Lin et al., 2025)) achieves under 19% accuracy, while the top informal model (Gemini-2.5-Pro) reaches almost 83% on the same problems—a 64-point gap on a matched problem set. Independent corroboration from Liang et al. (2025) shows informal accuracy exceeding 80% while formal success stays below 8% on the same competition problems. The gap is not noise; it reflects genuine differences in what formal verification and natural-language judgment are measuring.

LLM-as-judge vs. human grading. Dekoninck et al. (2025) provide the most direct comparison: on the Open Proof Corpus (>5,000 graded proofs), LLM-judge correlation with human expert grades reaches r=0.87 *with* structured rubrics and fine-tuning, but drops toward random without rubrics. Yang et al. (2025) report comparable correlations on ProofBench when frontier judges are paired with structured grading guidelines. Luong et al. (2025) achieve a higher r=0.93–0.96 with Gemini-2.5-Pro and detailed grading guidelines on IMO-ProofBench, but only 0.87 on a broader 170-system pool. Yue & Klein (2025) corroborate this: zero-shot LLM judging on out-of-distribution proof styles produces unreliable scores. JudgeBench (Tan et al., 2024) provides meta-evaluation infrastructure show-

Table 3. Representative 2024–2026 benchmarks classified by our five-axis taxonomy. FA = Final Answer; PR = Proof Required; F = Formal; NL = Natural Language; S = Static; L = Live.

Benchmark	Tier	Type	Form.	Contam.	Lang.	Scale
OlympiadBench (He et al., 2024)	Comp.	FA	NL	S	Multi	8,476
OmniMATH (Gao et al., 2024)	Comp.	FA	NL	S	EN	4,428
MathArena (Balunović et al., 2025)	Comp.	FA	NL	L	EN	Live
FrontierMath (Glazer et al., 2024)	Research	FA	NL	L	EN	Private
Humanity's Last Exam (Phan et al., 2025)	Research	FA	NL	S	EN	2,500
MiniF2F-v2 (Ospanov et al., 2025)	Comp.	PR	F (Lean/Isa)	S	EN	488
PutnamBench (Tsoukalas et al., 2024)	Undergrad	PR	F (Lean/Isa)	S	EN	657
FormalMATH (Yu et al., 2025)	Comp.+UG	PR	F (Lean)	S	EN	5,560
LeanDojo (Yang et al., 2023)	General	PR	F (Lean)	S	EN	122k+
RLMEval (Poiroux et al., 2025)	Research	PR	F (Lean)	S	EN	613
USAMO 2025 (Petrov et al., 2025)	Comp.	PR	NL	L	EN	6
Open Proof Corpus (Dekoninck et al., 2025)	Comp.	PR	NL	S	EN	5,000+
ProofBench (Yang et al., 2025)	Comp.	PR	NL	S	EN	145/435
IMO-ProofBench (Luong et al., 2025)	Comp.	PR	NL	S	EN	60
IMO-GradingBench (Luong et al., 2025)	Meta	PR	NL	S	EN	1,000
GHOSTS (Frieder et al., 2024)	Grad.	PR	NL	S	EN	709

ing that frontier LLM judges (including GPT-4 and Claude variants) fall well short of consistent agreement with reference answers on reasoning-heavy preference comparisons, reinforcing that LLM-judge reliability is far from settled even for frontier models. LLM-as-judge is not a drop-in replacement for human grading: the rubric and fine-tuning pipeline are load-bearing.

PRM step-level vs. final-answer accuracy. Zhang et al. (2025) document a systematic gap between PRM step-level AUC-PR and best-of-N selection accuracy. In their analysis, PRMs that achieve high step-level precision systematically fail to select the best final answer when the solution space contains *reward-hacking* trajectories that satisfy the step-level reward model while arriving at incorrect conclusions.

Quantified disagreement matrix. Table 4 consolidates the magnitudes of cross-paradigm divergence as reported in the literature, alongside the source from which each number is drawn.

6.2. A Small Confirmatory Study on OPC

Scope caveat. The study below covers Paradigms 3 (LLM-as-judge), 4 (PRM), and 5 (human grading); Paradigms 1, 2, and 6 are evidenced through the published-results synthesis in Section 6.1, not through new runs. We treat the study as supporting illustration, not as a complete six-paradigm calibration, and discuss the limits of this coverage in Section 9. A full six-paradigm matched study is left to the research agenda (Section 7).

Setup. We sampled 150 human-graded proofs from OPC's `generic` split (stratified: 75 human-correct, 75 human-incorrect, ties and uncertain rows dropped). Sample sizes in recent peer-reviewed proof-evaluation work range from 6 problems (Petrov et al., 2025) to 60 (Luong et al., 2025) to 435 (Yang et al., 2025); our n=150 falls within this range, and the OPC paper itself reports its primary judge evaluations on subsets of 112–293 proofs. The LLM-as-judge is Qwen2.5-7B-Instruct with a four-criterion structured rubric (logical validity, no gaps, correct conclusion, rigor). The PRM is Qwen2.5-Math-PRM-7B (Zhang et al., 2025) applied with the official `<extra_0>` step-separator protocol; we aggregate per-step rewards by minimum, geometric mean, and last-step at threshold 0.5. All models run on free Kaggle T4×2 (∼30 GB total GPU); we deliberately use open-source models rather than frontier closed-source judges to test whether the OPC findings generalize to the open-source regime.

Results. Table 5 reports Cohen's κ across paradigm pairs with bootstrap 95% CIs (1,000 resamples). All paradigm pairs exhibit only slight-to-fair agreement ($\kappa \in [0.04, 0.28]$). Only PRM under min-aggregation has a CI lower bound clearly above 0.10. The Human vs. LLM-Judge CI includes zero: we cannot statistically distinguish the open-source LLM-judge's agreement with humans from chance. Against the stratified ground-truth balance of 75/75, the LLM-judge labels 104 proofs correct (35% inflation) and the PRM under geometric mean labels 138 correct (84% inflation), indicating systematic over-acceptance by both automated paradigms.

Disagreement failure modes. Categorizing the 87 proofs (58%) where paradigms disagree under geometric-mean PRM aggregation, the dominant pattern (48.3%) is *correlated over-acceptance*: both automated paradigms accept a proof that humans reject. This is the most damaging pattern for naive ensembling, since the two automated signals share their failure. Under the better-aligned min-aggregation, this

Table 4. Cross-paradigm disagreement: quantified evidence from published results. Δ is the reported gap (or correlation) on matched problem sets.

Paradigm A	Paradigm B	Documented Δ
Formal verifier	NL judge	$\sim$19% vs. $\sim$83% on Putnam subset of OPC (Δ=64pp) (Dekoninck et al., 2025); <8% vs. >80% on competition problems (Liang et al., 2025)
LLM-as-judge	Human grading	r=0.87 w/ rubric, near-random without (Dekoninck et al., 2025; Yang et al., 2025; Yue & Klein, 2025); r=0.93–0.96 on IMO-ProofBench, 0.87 on 170-system pool (Luong et al., 2025)
PRM step score	Final-answer accuracy	High AUC-PR $\not\Rightarrow$ high BoN selection; reward hacking documented (Zhang et al., 2025)
Symbolic checker	LLM-as-judge	Non-trivial error rate on trig/piecewise answers (Dekoninck et al., 2025)
Open-source LLM-judge	Human (this work)	κ=0.13, 95% CI $[-0.01, +0.27]$ on OPC subset (Section 6.2)

drops to 21% but is replaced by 41% LLM-judge-only disagreement. Different aggregations surface different structural weaknesses.

Sensitivity to PRM aggregation. The same PRM model yields a 7$\times$ range in κ across aggregation rules (0.04 to 0.28). Qwen's recommended geometric mean is not the best aggregation for whole-proof judgment in our setting.

On the metric. Cohen's κ collapses the partial-credit grading scales used by human IMO graders (typically 0–7) into a binary correct/incorrect verdict, and assumes two raters per item. Intraclass correlation (ICC) or Krippendorff's α would be the natural extension for partial-credit and multi-rater settings; we anticipate that fine-grained agreement is likely *worse* than the binary κ reported here, since binary agreement absorbs much partial-credit variance. We retain κ for direct comparability with Dekoninck et al. (2025) and Tan et al. (2024), and flag the upgrade as part of the calibration agenda below.

Interpretation. Each paradigm operationalizes a different latent concept of mathematical correctness. Formal verifiers operationalize logical validity within a fixed type theory.

Table 5. Cross-paradigm Cohen's κ on 150 OPC proofs with bootstrap 95% CIs (1,000 resamples).

Pair	κ	95% CI
Human vs. LLM-Judge	0.133	$[-0.006, +0.271]$
Human vs. PRM (min)	**0.280**	$[+0.122, +0.436]$
Human vs. PRM (geo)	0.133	$[+0.047, +0.235]$
Human vs. PRM (last)	0.040	$[-0.037, +0.117]$
LLM-Judge vs. PRM (min)	0.160	$[+0.024, +0.298]$
LLM-Judge vs. PRM (geo)	0.131	$[+0.004, +0.272]$
LLM-Judge vs. PRM (last)	0.091	$[-0.029, +0.232]$

Human graders operationalize mathematical soundness plus insight and presentational clarity. PRMs operationalize local step plausibility, aggregated under whatever rule the deployer chooses. Symbolic checkers operationalize algebraic equivalence under a fixed normal form. None of these concepts subsumes the others, which is why paradigm agreement cannot be assumed. Our small study and the published evidence in Section 6.1 jointly support treating cross-paradigm calibration as a research-agenda item rather than an implementation detail.

7. Gap Analysis and Research Agenda

Our taxonomy and disagreement analysis reveal five structural gaps in the current evaluation ecosystem.

Gap 1: No verifier satisfies all desiderata simultaneously. Formal verifiers guarantee correctness but require autoformalization and are compute-intensive. LLM-as-judge scales but is sensitive to configuration. PRMs apply only to step-by-step natural-language settings. Human grading is the gold standard but is slow and expensive. No paradigm is simultaneously correct, scalable, and aligned with human mathematical judgment.

Gap 2: The formal/informal evaluation boundary is unmapped. A proof verified in Lean 4 is not necessarily a proof that a human mathematician would consider insightful or well-written; conversely, a proof graded highly by human experts may contain gaps undetectable without formal checking. Mapping the correlation between formal and informal evaluation on matched problem sets is a critical open question (Liang et al., 2025).

Gap 3: Meta-evaluation coverage is thin. JudgeBench (Tan et al., 2024) and IMO-GradingBench (Luong et al., 2025) provide initial meta-evaluation infrastructure for judge reliability, but coverage of proof-specific evaluation dimensions (gap detection, logical validity, rigor) remains limited.

Gap 4: Live benchmark infrastructure is immature. Static benchmarks are vulnerable to contamination as models train on ever-larger web corpora. MathArena (Balunović

et al., 2025) represents an early step toward live evaluation, but automated pipelines for generating, verifying, and scoring new problems at scale do not yet exist.

Gap 5: Generalization gaps—multilingual, cultural, and difficulty-calibrated. Most evaluation infrastructure targets English-language problems from Western mathematical traditions; OlympiadBench (He et al., 2024) is an exception, but systematic multilingual evaluation of *proof* quality remains an open problem. In parallel, benchmarks span a wide difficulty range from GSM8K to FrontierMath without an agreed-upon difficulty scale, so a model that scores 90% on MATH and 2% on FrontierMath occupies an ill-defined position relative to one that scores 80% on MATH and 5% on FrontierMath. Both are facets of the same problem: scores do not yet live on a common latent ability axis.

Research agenda. We propose four concrete directions: (1) *Calibration studies* that run all six verification paradigms on a shared matched problem set, producing empirical conversion factors between paradigm-specific scores. Our small study in Section 6.2 is a first step at three-paradigm scale; the full six-paradigm matched study is open. (2) *Unified proof datasets* providing parallel formal (Lean 4) and natural-language representations of the same problems, graded by both formal verifiers and human experts. (3) *Live evaluation infrastructure* combining autoformalization pipelines with human-in-the-loop problem generation and verification, making contamination-resistant benchmarking sustainable at scale. (4) *Difficulty calibration frameworks* mapping benchmark scores onto a common latent ability scale using item response theory (IRT) or similar psychometric methods.

8. The Integrated Evaluation Framework

Drawing on our taxonomy, we sketch a reference framework for evaluating AI-generated mathematical content. The framework operates in three stages and is modular: components can be adopted incrementally as infrastructure matures.

Stage 1: Problem classification. Classify each evaluation problem along two dimensions: (i) answer type (final-answer vs. proof-required) and (ii) formalization level (formal-ready vs. natural-language-only). Final-answer problems route to symbolic checkers (Paradigm 1) with an LLM-as-judge fallback for format robustness on problems where symbolic equivalence checking is known to be unreliable. Proof-required problems route to Stage 2.

Stage 2: Parallel verification. For proof problems, apply three verification signals in parallel and aggregate them:

1. *Formal verification* (Paradigm 2): attempt autoformalization to Lean 4 or Isabelle, then run the proof assistant. When autoformalization succeeds and the proof verifies, this signal is treated as a hard correctness certificate.
2. *LLM-as-judge* (Paradigm 3): apply a frontier judge model with a domain-specific structured rubric. Use multiple independent judge samples and report variance to quantify judge uncertainty.
3. *PRM step-level scoring* (Paradigm 4): for chain-of-thought outputs, apply a calibrated PRM to flag potentially incorrect reasoning steps. High step-level suspicion scores can trigger additional human review.

Aggregate the three signals with a learned ensemble calibrated against human expert ground truth. When signals conflict (e.g., the LLM judge scores a proof highly but the PRM flags several steps), flag the problem for human review rather than resolving automatically. Our confirmatory study in Section 6.2 shows that conflicts are common (58–61% of proofs) and that ensemble design must account for correlated over-acceptance between LLM-judge and PRM signals.

Stage 3: Human expert calibration. Periodically calibrate the ensemble against human expert grades on a held-out calibration set (e.g., annual IMO/USAMO entries). We recommend a minimum calibration set of 200 human-graded proofs per update cycle, consistent with sample sizes in Dekoninck et al. (2025) and Luong et al. (2025).

Failure modes and mitigations. Autoformalization is the primary bottleneck for Stage 2a: current success rates on olympiad problems are well below 50%, so the formal signal is often absent. Rubric design and judge fine-tuning are the bottlenecks for Stage 2b; our study shows open-source 7B judges with structured prompts reach only κ=0.13, motivating domain-specific fine-tuning rather than zero-shot judging for high-stakes use. Stage 3 calibration requires sustained expert involvement; community-organized annotation campaigns modelled on OPC and IMO-GradingBench are a natural path.

9. Discussion

Relation to existing integrated frameworks. The closest existing integrated evaluation work is IMO-Bench (Luong et al., 2025), which introduces IMO-AnswerBench, IMO-ProofBench, and IMO-GradingBench as a coordinated triad. IMO-Bench is a primary research paper rather than a survey; it instantiates one paradigm pair (LLM-as-judge calibrated against human experts) with strong empirical results (r=0.93–0.96 on ProofBench), but does not taxonomize the broader evaluation landscape or position the six verification paradigms against each other. Our contribution is complementary: we provide the methodological synthesis allowing

researchers to understand where any given evaluation approach sits in the broader ecosystem.

The contamination problem is worsening. As LLMs are trained on increasingly large web corpora, the probability that any static benchmark's problems appear in training data grows monotonically. The community has responded with contamination-resistant designs (FrontierMath's private problems, MathArena's live competition integration). We argue contamination resistance should be treated as a first-class design criterion alongside difficulty and coverage.

Limitations. The taxonomy is field-snapshot-dependent: new paradigms such as trained generative verifiers (Liu et al., 2025; Dekoninck et al., 2025) blur the LLM-judge/PRM boundary and will require updating. Our confirmatory study uses $n{=}150$ and open-source 7B models; frontier closed-source judges (Gemini-2.5-Pro at $r{=}0.96$ (Luong et al., 2025)) bracket the upper end of LLM-judge agreement, while our open-source result establishes a lower bound. The paradigm-divergence claim is robust across this range: even when the LLM-judge is moved from $\kappa{\approx}0.13$ to $r{\approx}0.96$ with humans, the formal/informal gap on matched problems remains >60 percentage points (Dekoninck et al., 2025; Liang et al., 2025). The study covers three of the six paradigms; the published-results synthesis in Section 6.1 covers the other three. A full six-paradigm matched calibration study, ideally on a shared problem set, remains the highest-leverage next experiment.

10. Conclusion

We have presented a paradigm-centric taxonomy of evaluation methodology for AI-generated mathematical solutions and proofs, crossing the formal/informal verification boundary. Our four primary contributions are: (1) a six-paradigm taxonomy of verification methodology organized by the correctness-guarantee vs. scalability trade-off; (2) a five-axis classification of over thirty 2024–2026 benchmarks; (3) direct evidence—a published-results synthesis and a small confirmatory study—that paradigms diverge systematically on matched problem sets; and (4) a gap analysis identifying five structural deficiencies, paired with a four-direction research agenda.

A central finding is that no existing verification paradigm simultaneously satisfies correctness guarantees, scalability, and alignment with human mathematical judgment. The benchmark landscape has shifted from saturated final-answer benchmarks to far-from-saturated natural-language proof benchmarks ($<25\%$ on USAMO (Petrov et al., 2025)) and research-level formal benchmarks (10% on RLMEval (Poiroux et al., 2025)). The highest-leverage near-term contribution remains a calibration study running all six paradigms on a shared matched problem set. As AI systems approach human performance on mathematical reasoning, the rigor of how we measure that performance becomes as important as the performance itself.

Impact Statement

This paper presents a survey and taxonomy of evaluation methods for AI-generated mathematical content, supported by a small confirmatory study. The primary societal implication is enabling more reliable and trustworthy assessment of AI mathematical capability, which matters for both AI safety (understanding actual model capabilities) and educational applications. We do not introduce new models or training procedures and see no significant negative societal consequences specific to this methodological contribution.

References

Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., and Yin, W. Large language models for mathematical reasoning: Progresses and challenges. *EACL 2024 SRW; arXiv:2402.00157*, 2024.

Amjad, H., Shahzad, R. K., Shahzad, A., and Fatima, M. Mathematical reasoning in large language models: Benchmarks, architectures, evaluation, and open challenges. *arXiv:2605.19723*, 2026.

Balunović, M., Dekoninck, J., Petrov, I., Jovanović, N., and Vechev, M. Matharena: Evaluating llms on uncontaminated math competitions. *matharena.ai; technical report*, 2025.

Dekoninck, J., Petrov, I., Minchev, K., Balunović, M., Vechev, M., et al. The open proof corpus: A large-scale study of llm-generated mathematical proofs. *ICLR 2026 (Poster); arXiv:2506.21621*, 2025.

Fourrier, C., Habib, N., Wolf, T., and Tunstall, L. LightEval: A lightweight framework for LLM evaluation. https://github.com/huggingface/lighteval, 2024. Hugging Face.

Frieder, S., Pinchetti, L., et al. Mathematical capabilities of chatgpt (ghosts dataset). *NeurIPS 2023 Datasets; arXiv:2301.13867*, 2024.

Gao, B., Song, F., Yang, Z., Cai, Z., Miao, Y., Dong, Q., Li, L., Ma, C., Chen, L., Xu, R., et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv:2410.07985*, 2024.

Glazer, E., Erdil, E., Besiroglu, T., et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv:2411.04872*, 2024.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *ACL 2024; arXiv:2402.14008*, 2024.

Li, Z. et al. A survey on deep learning for theorem proving. *COLM 2024; arXiv:2404.09939*, 2024.

Liang, Z., Song, L., Li, Y., Yang, T., Zhang, F., Mi, H., and Yu, D. Towards solving more challenging imo problems via decoupled reasoning and proving. *arXiv:2507.06804*, 2025.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *ICLR 2024; arXiv:2305.20050*, 2023.

Lin, Y., Tang, S., Lyu, B., Yang, Z., Chung, J.-H., Zhao, H., Jiang, L., Geng, Y., Ge, J., Sun, J., et al. Goedel-prover-v2: Scaling formal theorem proving with scaffolded data synthesis and self-correction. *arXiv:2508.03613*, 2025.

Liu et al. Scaling generative verifiers for natural language mathematical proof verification and selection. *NeurIPS 2025; arXiv:2511.13027*, 2025.

Luong, T., Hwang, D., Nguyen, H. H., Ghiasi, G., Chervonyi, Y., et al. Towards robust mathematical reasoning. *EMNLP 2025 (main); arXiv:2511.01846*, 2025.

Meurer, A., Smith, C. P., Paprocki, M., Čertík, O., Kirpichev, S. B., Rocklin, M., Kumar, A., Ivanov, S., Moore, J. K., Singh, S., Rathnayake, T., Vig, S., Granger, B. E., Muller, R. P., Bonazzi, F., Gupta, H., Vats, S., Johansson, F., Pedregosa, F., Curry, M. J., Terrel, A. R., Roučka, v., Saboo, A., Fernando, I., Kulal, S., Cimrman, R., and Scopatz, A. Sympy: Symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017. doi: 10.7717/peerj-cs.103.

Ospanov, A. et al. minif2f-lean revisited: Reviewing limitations and charting a path forward. *arXiv:2511.03108*, 2025.

Petrov, I., Dekoninck, J., Baltadzhiev, L., Drencheva, M., Minchev, K., Balunović, M., Jovanović, N., and Vechev, M. Proof or bluff? evaluating llms on 2025 usa math olympiad. *arXiv:2503.21934*, 2025.

Phan, L. et al. Humanity's last exam. *arXiv:2501.14249*, 2025.

Poiroux, Q. et al. Rlmeval: Evaluating research-level neural theorem proving. *arXiv:2509.15700*, 2025.

Ren, Z. Z., Shao, Z., Song, J., et al. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv:2504.21801*, 2025.

Tan, S. et al. Judgebench: A benchmark for evaluating llm-based judges. *arXiv:2410.12784*, 2024.

Tsoukalas, G. et al. Putnambench: A multilingual competition-mathematics benchmark for formal theorem-proving. *NeurIPS 2024 Datasets; arXiv:2407.11214*, 2024.

Wang, P., Li, L., Shao, Z., Xu, R., Dai, D., Li, Y., Chen, D., Wu, Y., and Sui, Z. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *ACL 2024; arXiv:2312.08935*, 2024.

Wang, P. et al. A survey on large language models for mathematical reasoning. *ACM Computing Surveys*, 2025.

Wei et al. A survey on process reward models for llm reasoning. *arXiv:2502.10571*, 2025.

Yan, Y. et al. A survey of mathematical reasoning in the era of multimodal large language models. *ACL 2025; arXiv:2412.11936*, 2025.

Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., and Anandkumar, A. Leandojo: Theorem proving with retrieval-augmented language models. *NeurIPS 2023; arXiv:2306.15626*, 2023.

Yang, W. et al. Reliable fine-grained evaluation of natural language math proofs. *ICLR 2026; arXiv:2510.13888*, 2025.

Yu, Z. et al. Formalmath: Benchmarking formal mathematical reasoning of large language models. *arXiv:2505.02735*, 2025.

Yue, J. and Klein, D. Benchmarking llms on advanced mathematical reasoning. *UC Berkeley M.S. Thesis; EECS-2025-121*, 2025.

Zhang, Z., Zheng, C., Wu, Y., Zhang, B., Lin, R., Yu, B., Liu, D., Zhou, J., and Lin, J. The lessons of developing process reward models in mathematical reasoning. *arXiv:2501.07301*, 2025.

Zheng et al. A survey of process reward models: Foundations, approaches, and future directions. *arXiv preprint*, 2025.