Faster Than They Can Be Checked

Measuring the verification gap in AI-assisted solutions to Erdős problems

Preprint. Subomi Olagoke.

Abstract

Terence Tao has predicted that "AI-generated proofs will accumulate faster than they can be verified." This paper measures that prediction on the largest public record of AI-assisted research mathematics: the proof claims submitted to erdosproblems.com. It combines four open sources:

- all 251 proof claims posted between 14 July and 24 September 2026, each with its full comment thread, hand-coded for verdicts;
- 2,614 status changes from the community database's git history;
- the edit history of Tao's ledger of AI contributions;
- 2,112 logged AI attempts from a community attempt archive.

Of the 224 claims on problems that were open when claimed:

- 97% were produced with AI;
- 48% received any comment;
- 19% reached any verdict;
- 14% were confirmed as correct, fully or partially;
- 4% were accepted by the site's moderators.

Verdicts arrive quickly or not at all. The median verdict came 1.1 days after posting, yet a Kaplan–Meier estimate leaves 82% of claims with no verdict after 60 days. When data collection ended, 185 claims were waiting, with a median wait of 53 days.

Attaching a Lean formalization did not raise the verdict rate (17% with Lean against 20% without). One claim presented as formally verified contained a custom axiom that assumed its key step. The verdicts that did happen came from 19 people, and three of them gave more than half.

We also record a mistaken dismissal: a complete, formally verified solution was waved off as "already solved", and nobody corrected it. We argue that the scarce resource in AI-assisted

mathematics is expert adjudication, not proof generation or proof checking. We release the dataset and code.

1. Introduction

Paul Erdős left behind more than a thousand problems, catalogued since 2023 at erdosproblems.com [Bloom]. In 2025–26 they became the most visible test of whether AI systems can do research mathematics:

- Hobbyists and undergraduates using public models produced a stream of results [Quanta].
- Laboratories published case studies [Feng].
- A controlled benchmark followed [FME].

Tao started a public ledger of AI contributions and maintained it until 30 June 2026 [Tao-wiki].

Solving is only half of the process. A claimed proof counts as mathematics only when someone qualified has checked it, placed it against the literature, and vouched for it. Tao's 2026 ICM lecture anticipates the imbalance [Tao-ICM]:

> "Sites devoted to collecting mathematical problems, such as the Erdős problems database, already contain dozens of AI-generated proof submissions. Many of these are likely to be correct; but in a substantial number of cases no human expert has yet volunteered to verify and vouch for them."

He predicts that "AI-generated proofs will accumulate faster than they can be verified", and a transition "from an era of proof scarcity to an era of proof abundance." The prediction is qualitative. As far as we can find, nobody has measured it (§2).

This paper measures it. We treat every public proof claim on erdosproblems.com as an observation and ask five questions:

1. How many claims arrive?
2. What fraction ever receive a verdict, and how quickly?
3. What are the verdicts?
4. Which factors predict a verdict?
5. Who delivers the verdicts?

The design follows a principle from our earlier work on change detection [Olagoke]: to measure how often a system is right, measure it where the answer is not in doubt, and report what comes

back. Here the "system" is the community process that turns claims into mathematics.

Contributions.

1. **The first measurement of verification debt in AI-assisted mathematics.** We quantify the claim-to-verdict funnel, how long verdicts take (with survival analysis), and the growing backlog.
2. **A hand-coded dataset of 251 proof claims**, each with its verdict, the time of the verdict, who gave it, and a verbatim supporting quote, released with the collection and analysis code.
3. **Findings that bear on policy.** A Lean formalization does not predict scrutiny. Verification rests on very few people. There is evidence of errors in both directions: flawed claims and wrongly dismissed ones.

2. Related work

Laboratory case studies. Feng et al. ran Gemini on 700 open Erdős problems [Feng]. Of 200 candidate solutions they checked, "137 (68.5%) of responses were fundamentally flawed, while 63 (31.5%) of responses were technically correct, but only 13 (6.5%) were meaningfully correct." Eight of those 13 turned out to be in the literature already. They report that "the most challenging step for human experts was not verification, but determining if the solutions already existed in the literature." That study measures one pipeline's internal outputs before release. We measure public claims after release.

Benchmarks. FrontierMath Erdős tests five models on 68 curated open problems, with Lean proofs required [FME]. GPT-6 Astra solved 2 of the 68, and the other four models solved none. It is a controlled measure of capability on hard problems. Ours is an observational study of the whole flow of claims, most of which concern easier problems.

Verification economics. Kallel and El Louadi argue that "machine checking produces verification abundance while leaving adjudication scarce" [Kallel]. That is a theoretical claim, and our data tests it. The term "verification debt" comes from software engineering, where it describes AI-generated code outrunning review [Sonar]. We borrow it. The nearest quantitative parallels are the curl bug-bounty programme, where about 5% of 2025 submissions were real vulnerabilities and each report occupied 3–4 people [curl], and journal review, where AI-heavy submissions pass through a much narrower funnel [OrgSci].

Reporting bias. Tao warns that "successes are announced and failures are not" [Tao-ICM], and his ledger notes that reports of negative results are "most likely incomplete" [Tao-wiki]. The forum's proof-claims pages differ, because a claim is recorded before anyone judges it. That makes them the least outcome-selected public record we know of. They are still selected, though: a claimant

decides to post. The attempt archive of Ivanisvili and Seven [Hunter] records failed attempts too, and we use it to see what happens further upstream.

3. Data and methods

Sources. All data are public and were collected on 25 September 2026.

Source	Content	Size
erdosproblems.com problem pages	status label, comment count, proof-claim count	1,221 problems
Proof-claims pages and threads	type (full or partial), AI tools, submission time, links to proof and Lean formalization, comment thread	251 claims (IDs 2–348)
teorth/erdosproblems git history	every change to a problem's status	1,366 commits, 2,614 changes
Tao's AI-contributions ledger	categorized AI contributions with outcome colours, and its edit history	538 rows, 1,086 edit events
LLM Hunter archive [Hunter]	AI attempt write-ups and community reviews	2,112 attempts, 22 reviews

The site assigns claim IDs in sequence. Of IDs 2–348, 97 do not appear on any problem page. We take these to be removed or unpublished claims. We report this but cannot observe them.

Status at the time of the claim. For each claim, we replay the database's git history to find the problem's status at the moment of submission. A claim counts as "on an open problem" unless the problem was already marked proved, disproved or solved. By this rule, 224 of the 251 claims were on open problems. The other 27 concern problems already solved.

Verdict coding. Every claim that drew at least one comment (118 claims) was read in full and assigned one verdict. The coding was done by an AI model, following these written rules (see Use of AI):

- `verified_correct`, `verified_partial`: a credible reviewer states the proof is correct, or a moderator accepted it;
- `refuted`: a concrete error was found and not repaired;
- `already_known`: the result is shown to be in the literature;
- `misstated_or_technicality`: the claim solves a different or trivial version of the problem;

- `superseded`: another claim got the credit;
- `dismissed_in_error`: a reviewer's dismissal is contradicted by the site's own record;
- `discussed_no_verdict`, `other`: comments without a verdict, or only the claimant's own messages.

Each verdict also records the date of the comment that settled it, who gave it, and a verbatim quote of at most 30 words, which was checked by script against the thread.

These do not count as verdicts: automated screens ("GPT finds no issues"), relayed AI reviews, and bare approvals.

A second pass audited a random sample of 10 claims plus every `already_known` and borderline code, and changed one code (§5.1). The 133 claims with no comments are coded `no_comments`. Moderator acceptance appears on the site as a green highlight on the claim, and we checked it against the database: for all seven accepted full claims, the problem changed status within days. The two accepted partial claims leave the status unchanged, as expected.

Time to verdict. Latency runs from submission to the settling comment. Claims without a verdict are censored at the collection date, and we estimate the survival function with Kaplan–Meier. Three verdicts come from credits on problem pages and have no date. They count toward verdict rates but not toward latency.

The ledger. It was restructured three times (December 2025, January–February 2026, April 2026). We therefore follow each problem across versions rather than following sections.

Ethics. All data are public posts on an open forum and public repositories, and no private data were used. Volunteer reviewers and claimants are not named in this paper. The only person named is the site's moderator, whose reviewing role is public. In the released dataset, forum usernames are replaced by stable pseudonyms, full comment texts are omitted, and each claim links to its public thread.

The attempt archive. We count the attempt files for Erdős problems and the review records, and take the review labels as given.

4. Results

4.1 The funnel

Among the 224 claims on open problems (Fig. A):

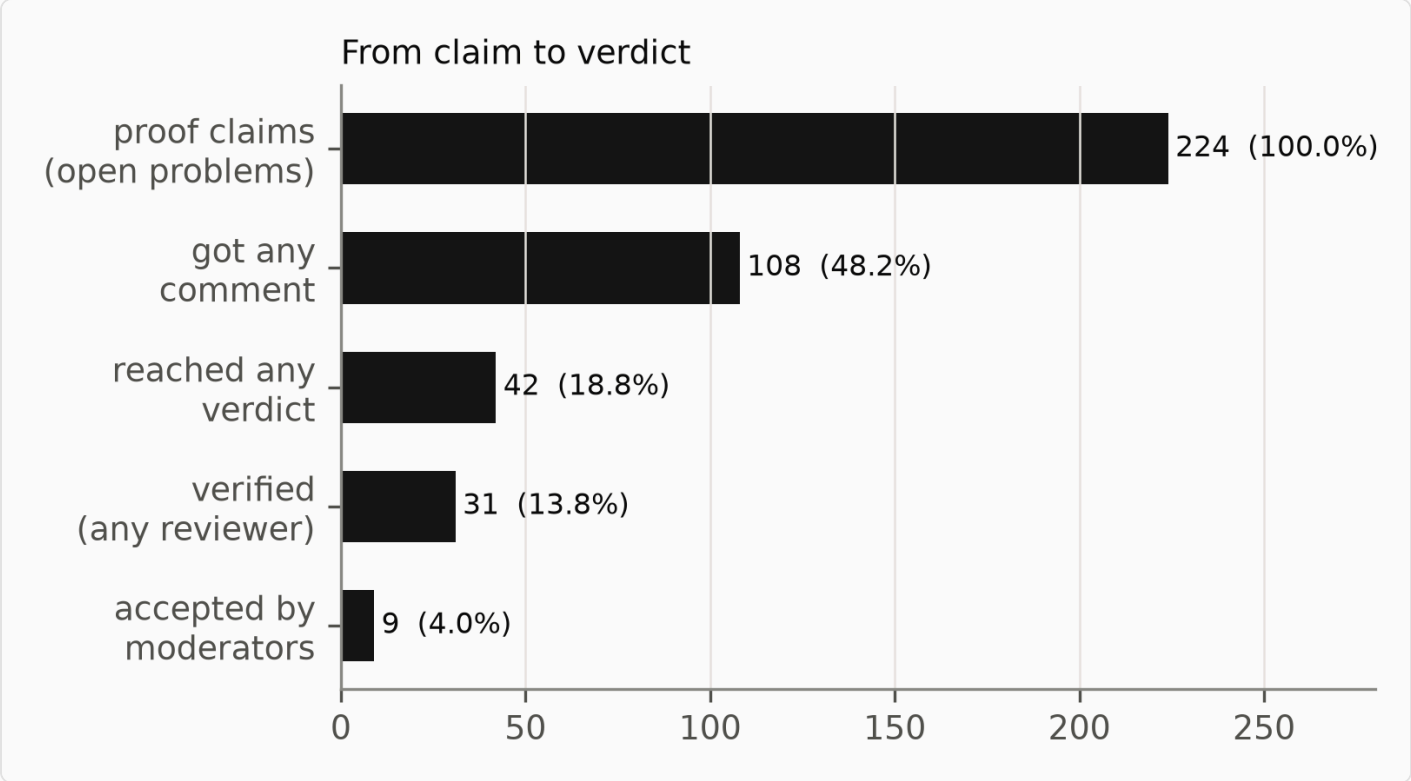
Stage	Claims	Share
Posted	224	100%
Received any comment	108	48.2%
Reached any verdict	42	18.8%
Verified (fully or partially)	31	13.8%
Accepted by moderators	9	4.0%

Almost all claims were AI-assisted: 244 of all 251 (97.2%) name an AI tool. Among the 116 full-solution claims on open problems, 54 (46.6%) received no comment at all.

The verdicts are overwhelmingly positive:

Verdict	Count
Verified correct or partial	31
Already in the literature	6
Misstated or technicality	3
Refuted	1
Dismissed in error	1

This is consistent with claims being positively selected before they are posted. It is also consistent with reviewers engaging mainly with claims that look promising. Silence, not refutation, is the typical outcome.



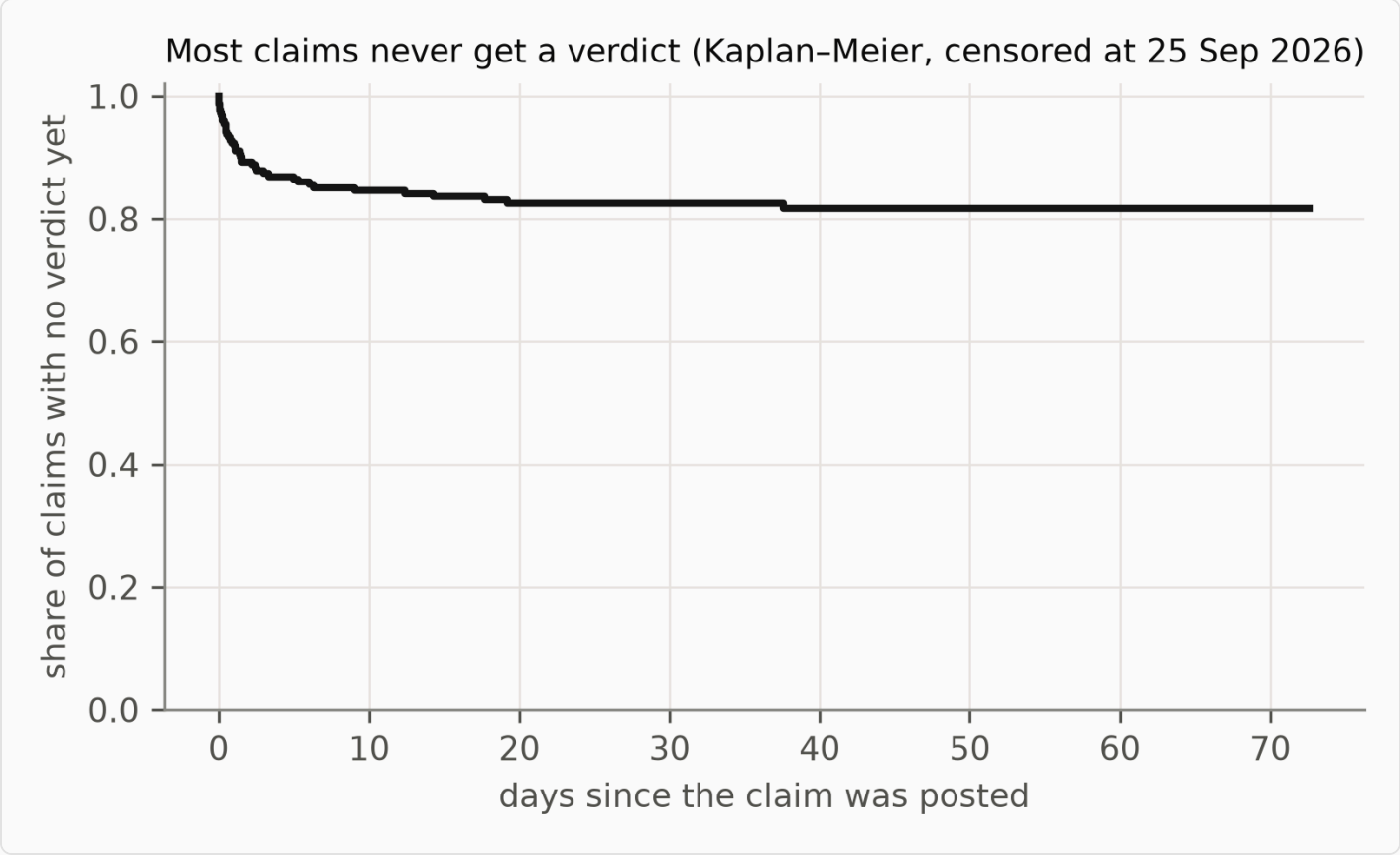
From claim to verdict: 224 proof claims on problems open when claimed, 14 Jul – 24 Sep 2026.

4.2 Verdicts come quickly or not at all

Among claims that reached a dated verdict (n = 39), the median time to verdict was **1.1 days** and the 75th percentile was 5 days. The Kaplan–Meier estimate of the share of claims still without a verdict is:

Days after posting	Share with no verdict
7	0.85
30	0.82
60	0.82

After the first week, the curve is essentially flat (Fig. B). A claim that is not looked at within days is, in practice, never looked at. At the collection date, 185 claims on open problems had no verdict, and the median time they had been waiting was 53 days.

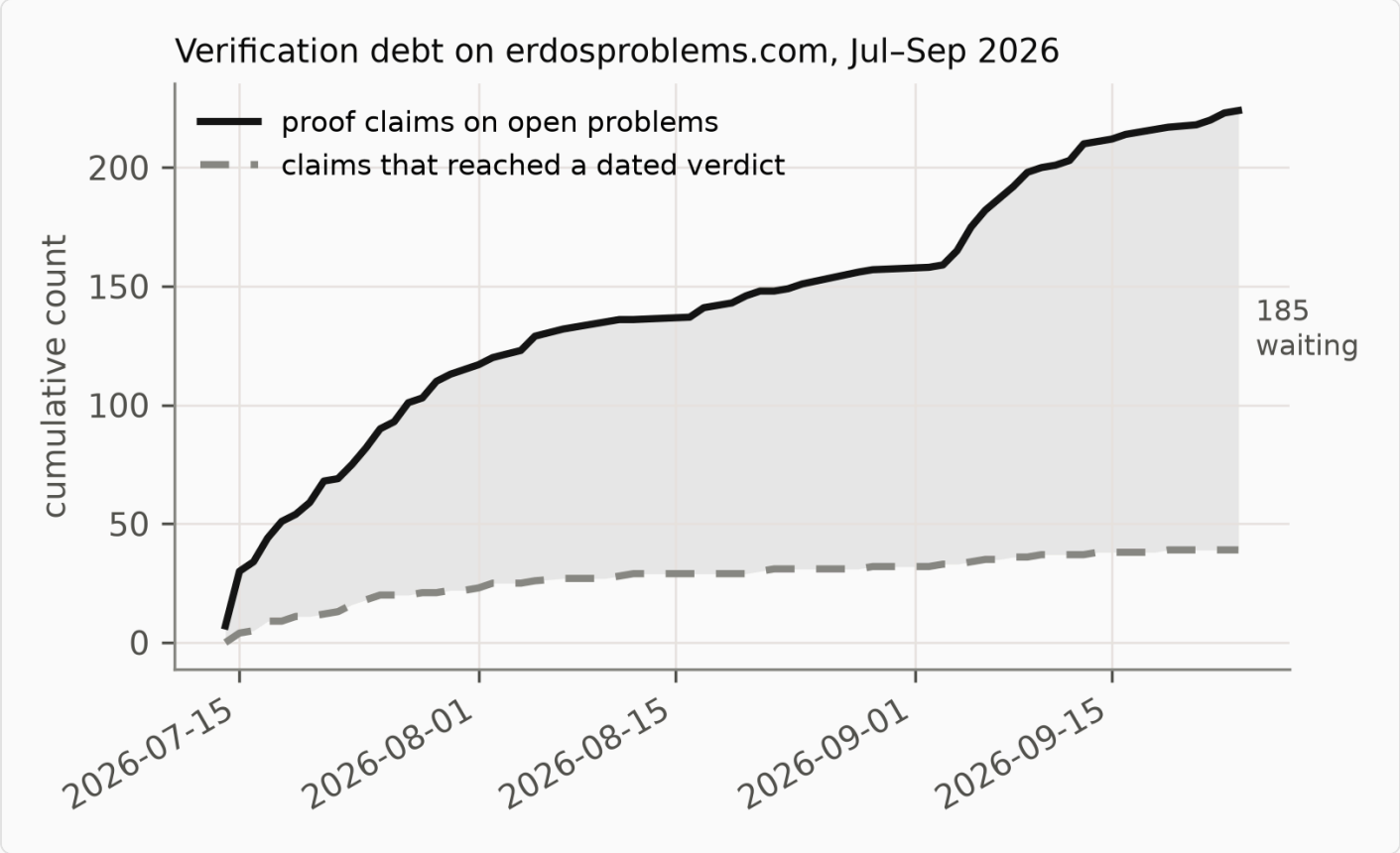


Share of claims with no verdict against days since posting (Kaplan–Meier, censored at 25 Sep 2026).

4.3 The backlog

Claims arrived at about 120 in July, 50 in August and 80 in September. Dated verdicts accumulated to 39 (Fig. C). The gap between the two curves, the verification debt, grew steadily over the whole period and never shrank.

For scale, the database records 187 problems newly marked solved between September 2025 and September 2026, counting each problem once and excluding the initial import. Over 2024 to August 2025, Quanta reports 111 problems moving from open to solved [Quanta]. The public record of AI proof claims now outpaces the community's rate of recorded resolutions.



Cumulative proof claims on open problems, and claims with a dated verdict. The shaded area is the verification debt.

4.4 What predicts a verdict

Group	n	Any verdict	Verified	No comment
With Lean formalization	106	17.0%	11.3%	51.9%
Without Lean	118	20.3%	16.1%	51.7%
Full claim	116	22.4%	15.5%	46.6%
Partial claim	108	14.8%	12.0%	57.4%
GPT-5.6 family	123	23.6%	18.7%	45.5%
GPT-6 Astra	48	8.3%	4.2%	62.5%

Lean does not attract scrutiny. A formalization should make checking cheap, yet claims that came with one were no more likely to receive a verdict, and were slightly less likely to be verified. Of the 118 commented threads, only 47 discussed the formalization at all. Several verdicts rest on a reviewer rebuilding the Lean project rather than reading the argument. We report these separately, as a weaker kind of verification.

A formalization's value also depends on its statement and its axioms. In claim 172, a commenter pointed out that the formal development declares a custom `axiom growth_ineq`, so the key inequality was assumed rather than proved. Nobody followed up. This agrees with Firsching et al., who document 291 fixed cases of research-level misformalization [FC], and with Kallel and El Louadi's point that proof checking moves the burden onto adjudicating what the statement says [Kallel].

Newer models, less scrutiny. Claims made with GPT-6 Astra, the most recent system, reached a verdict least often. Our data cannot tell whether this reflects lower quality, larger volume, or reviewer fatigue.

By field, claims tagged "additive combinatorics" (n = 13) and "graph theory" (n = 40) reached verdicts most often, at 38.5% and 35.0%. "Analysis" (n = 29) reached them least often, at 6.9%. These groups are small.

4.5 Who checks

The 47 verdicts in the coded data (including those on already-solved problems) came from **19 people**. The top three gave 53.2%: moderator Thomas Bloom (10), a single non-moderator user (10), and one other user (5). Moderators gave 12 verdicts in total.

Adjudication of AI-assisted Erdős claims therefore rests on a handful of volunteers. For comparison, the attempt archive holds **2,112 AI attempts and 22 reviews**. Of those reviews, 12 found the result already known, 8 found a technicality, 1 found an error, and 1 accepted the attempt [Hunter].

4.6 The ledger

When Tao's ledger was frozen on 30 June 2026, 43 of its 256 primary entries (16.8%) still carried the "unverified candidate" mark. Of the 18 problems whose first entry was unverified, one received a verdict before the freeze. The ledger's own timeline shows the same pattern as the forum: entries were added faster than they were resolved.

5. Case studies

5.1 A correct proof, dismissed in error (#742)

Problem #742 asks whether every diameter-2-critical graph on n vertices has at most $\lfloor n^2/4 \rfloor$ edges. Füredi proved this for all sufficiently large n [Füredi], but the threshold is "a gigantic number: roughly a tower of 2's of height 10^{14}" [DFH]. The site therefore lists the problem as *decidable* ("resolved up to a finite check"), and Tao commented that it "remains open."

On 5 August 2026, a claimant submitted a complete, AI-generated proof for all n, with a Lean formalization, and said openly that "I don't have the background to validate the proof myself." The only reply, the same day, was: "This was already solved." That is true only for large n. Nobody corrected it. Apart from the claimant's own reply five days later, nothing followed, and at collection time, seven weeks on, the claim had no further response.

We checked the formalization independently:

- built unmodified at the pinned commit;
- the problem's official formal statement [FC-742] restated word for word in a separate file, and proved using only the claim's final theorem;
- only the three standard axioms used;
- the proof replayed through `leanchecker`.

On that evidence, the formal-conjectures version of #742 is proved. The episode shows that errors run both ways. Unchecked claims can be wrong, and claims can also be rejected wrongly, by a verdict that nobody reviews.

5.2 A formalization that assumes its key step (claim 172)

See §4.4. A "Lean-verified" label on a claim gives no guarantee unless someone checks its axioms and statement.

6. Limitations

- **Selection.** We observe only claims that someone chose to post. The 97 missing IDs cannot be observed. Claims on already-solved problems (27) are left out of the main funnel.
- **Verdicts outside the forum.** A claim may have been checked by email, in a preprint, or in another forum, and we would not see it. Our verdict rates are therefore lower bounds on scrutiny. They are not lower bounds on correctness.
- **Coding.** An AI model coded all threads under written rules, and a second pass audited a sample. The coding has not been replicated by a human coder. The released codebook, quotes and thread links

are designed to make that easy. Borderline rules, such as whether a Lean rebuild counts as verification, are stated in the released coding notes, and the rates change little under the alternatives.

- **Timing.** September claims have had less time to receive a verdict. Survival analysis accounts for this, but the tail of the curve is still thin.
- **Scope.** Erdős problems skew towards problems that are easy for AI [Tao-Mastodon, FME]. Rates for harder problems may differ.

7. Discussion and recommendations

The data support Tao's prediction and sharpen it. The bottleneck is not *checking* in the mechanical sense: half the claims arrive with machine-checkable proofs. The bottleneck is **adjudication**:

- deciding what a formal statement says;
- placing a result in the literature;
- being willing to vouch for it in public.

Only a few people supply this, and it does not grow with AI-generated output. We suggest four measures:

1. **Track claims to a verdict.** Give each claim a visible status (unreviewed, under review, verified, refuted, known), with a timestamp. A claim that goes silent should look different from one that has been judged.
2. **Standardize formal-verification checklists.** Require the statement to come from a third party such as formal-conjectures, publish the axiom output, and replay the proof in the kernel. Claims that meet the checklist can then be triaged cheaply, and claims that declare their own axioms flagged.
3. **Review dismissals, not just claims.** A one-line "already solved" should cite the result it relies on.
4. **Share the literature search.** The costly step, as Feng et al. also found, is knowing whether something is new. Literature checks on a per-problem basis, done once and shared, would help everyone who submits.

Data and code. The claims dataset (with coded verdicts and quotes), the status timelines, the ledger events, and all analysis code are released at https://github.com/ubmids/erdos-verification-gap. The licensing is mixed and set out in the repository's README: our annotations under CC BY 4.0, the code under MIT, and data derived from the Erdős problems database under Apache 2.0.

Use of AI

An AI model (Claude, Anthropic) assisted with code, literature search, verdict coding of claim threads (under the rules in the codebook, with every verdict supported by a script-verified verbatim

quotation), and drafting. The author directed the work, reviewed all outputs, and takes full responsibility for the content.

References

- [Bloom] T. F. Bloom, *Erdős Problems*, https://www.erdosproblems.com (accessed 25 Sep 2026).
- [Tao-wiki] T. Tao et al., "AI contributions to Erdős problems", github.com/teorth/erdosproblems/wiki (frozen 30 Jun 2026).
- [Tao-ICM] T. Tao, "Mathematics in the age of AI", arXiv:2608.16753 (2026).
- [Tao-Mastodon] T. Tao, Mathstodon post, 17 Jan 2026, https://mathstodon.xyz/@tao/115911902186528812.
- [Feng] T. Feng, T. Trinh, G. Bingham, et al., "Semi-Autonomous Mathematics Discovery with Gemini: A Case Study on the Erdős Problems", arXiv:2601.22401 (2026).
- [FME] T. Adamczewski and T. F. Bloom, "FrontierMath Erdős", arXiv:2609.25050 (2026).
- [Kallel] M. Kallel and M. El Louadi, "Verification abundance, adjudication scarcity: what happens to mathematical knowledge when proof checking becomes free", arXiv:2608.28997 (2026).
- [FC] M. Firsching, P. Lezeau, S. Mercuri, et al., "Formal Conjectures: An Open and Evolving Benchmark for Verified Discovery in Mathematics", arXiv:2605.13171 (2026).
- [DFH] A. Dailly, F. Foucaud and A. Hansberg, "Strengthening the Murty–Simon conjecture on diameter 2 critical graphs", *Discrete Math.* **342**(11) (2019), 3142–3159. arXiv:1812.08420.
- [Füredi] Z. Füredi, "The maximum number of edges in a minimal graph of diameter 2", *J. Graph Theory* **16** (1992), 81–98.
- [FC-742] google-deepmind/formal-conjectures, `FormalConjectures/ErdosProblems/742.lean`.
- [Hunter] P. Ivanisvili and M. Seven, *Erdős Problems LLM Hunter*, github.com/mehmetmars7/Erdosproblems-llm-hunter (accessed 25 Sep 2026).
- [Quanta] K. Kakaes, "Why the Legendary Erdős Problems Are Falling to AI", Quanta Magazine, 3 Aug 2026.
- [Sonar] Sonar, "AI code verification debt", sonarsource.com (accessed 2026).
- [curl] D. Stenberg, "Death by a thousand slops", daniel.haxx.se, 14 Jul 2025.
- [OrgSci] Organization Science editors, "More versus better, part I", orgsci.substack.com.
- [Olagoke] S. Olagoke, "What Does a Change Detector Find Where Nothing Changed?", SSRN preprint (2026).