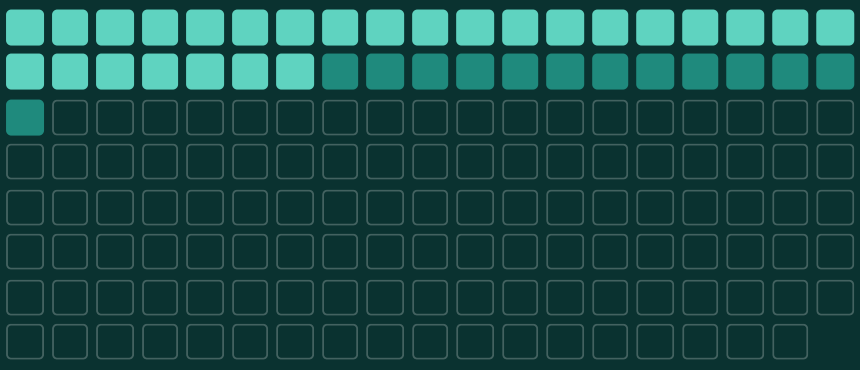
Papex

PAPEX REPORT NO. 1

Who checks AI-made proofs?

A count of 151 proof claims on Erdős problems



151 proof claims that name an AI system. For 39 of them we found a public check by someone else within 45 days.

Ahmet Tahir Mat, Papex Labs
ORCID 0009-0001-6664-6600

November 2026
Version 1 · https://doi.org/10.5281/zenodo.24378081

39 of 151

claims with a public check by someone other than the claimant within 45 days (25.8%)

26 of 151

claims where the proof itself was examined (17.2%)

32 to **66**

the same count with every tested reading of the rule that lowers it, or raises it, taken together

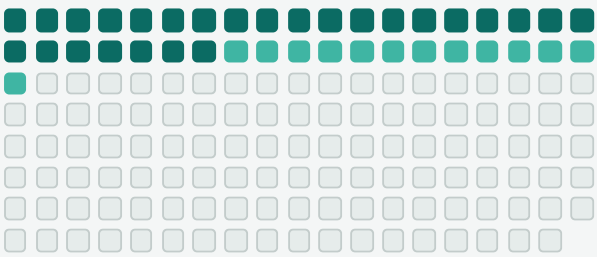
Summary

Between 14 July and 21 August 2026, 151 proof claims that name an AI system were submitted to erdosproblems.com and are still listed there. We read the public pages of each one in a fixed set of places and asked a single question: within 45 days, did anyone other than the claimant publicly check it?

For 39 of the 151 (25.8%; 95% interval 19.5% to 33.3%) we found at least one such check. For 26 (17.2%) at least one check was of the proof itself: someone read it and said what they had checked, formalised the result independently, or wrote the argument out step by step. For the other 13 the formal statement was reviewed or the files were run again.

The number depends more on what is allowed to count as a check than on statistical uncertainty. Taking together every tested reading of our rule that lowers it gives 32; every one that raises it, 66.

This is a count on one site, over one summer, in one field. "We found no check" does not mean "nobody checked".



26 The proof itself was examined: read, formalised independently or written out step by step.

13 Only the formal statement was reviewed or the files were run again.

112 No counted check found within 45 days.

Figure 1. Each square is one of the 151 claims that name an AI system, by what we found within 45 days.

1 The question

On the site counted here, 270 of the 290 listed proof claims name an AI system, and its owner froze new proof claims and problem comments on 6 October 2026, writing in the same post that people now mainly used the site to advertise their AI-generated proofs, "often without any attempt to explain them". Whether such proofs are right is one question; who looks at them is another. In an earlier, short search of the literature (not a systematic review) we did not find a measurement of how often results made with AI are checked by someone other than the person who claims them, or by whom. This note reports a count for one place where such claims are made in public.

2 What was counted

The claims. The site opened a feature for submitting proof claims on 14 July 2026 and froze new proof claims and problem comments on 6 October 2026. On 19 November 2026, 290 claims from that period were listed (deleted claims are not); the titles of 270 name an AI system, 10 do not, and 10 are unclear. We took every claim submitted from 14 July to 21 August, so that each had a full 45 days on the site before it froze. That is 154 claims: 151 that name an AI system and 3 that do not. All figures below are for the 151. The 3 are too few to compare with anything; we found no counted check for any of them.

We had first meant to compare claims made with and without AI. The site does not allow it: almost every proof claim there names an AI system.

A counted check. A public, dated record, written by someone who is not on the claimant's side (the claimant, co-authors and the account that submitted the claim), that belongs to this claim and is one of six kinds:

- **Reading.** The person says they read the proof or a named part of it and says whether it holds, or points to a specific error, gap or fix.
- **Independent formalisation.** The person formalises the result in a proof assistant and makes it public, or states that it was accepted without gaps.
- **Statement review.** The person checks that the formal statements the claim's formal proof is attached to say what the problem asks, or shows a mismatch.
- **Referee decision.** A record that the linked paper was accepted by a journal.
- **Re-run.** The person runs the claim's code or formal files and writes down the result.
- **Exposition.** The person writes out the argument step by step in their own words.

An approval that does not say what was read ("looks right", or "LGTM", short for "looks good to me") is not counted. Neither is a remark about novelty, an objection to how much of the problem the claim answers, or a model's verdict passed on without the person's own judgement. These were recorded; section 4 shows what happens when approvals, relayed model verdicts and scope objections that point to a specific part are counted.

Where we looked. The claim's comments; the problem's discussion thread; the problem's notes and expositions; pull requests in the Formal Conjectures and erdosproblems repositories on GitHub; the journal reference on arXiv; and the site's three "AI Contributions" threads. Which pages to open for a claim was decided by a script from fixed rules, not by the coder. A page beyond the list could be opened only through the script and with a written reason; in the final data 7 such pages were opened. Pages were fetched one request at a time, without signing in to any service; a page already saved was not fetched again, and a few failed requests were tried a second time.

The clock. It starts on the day the claim was submitted, or earlier if the same account had already posted the same text or document on the site. The main measure is 45 days, the 45th day included. Intervals are 95% Wilson intervals.

The rule came before the count. The protocol was written and saved before any new case was opened; it followed a 30-case pilot, whose cases were coded again. Coders were told not to make rules and not to add up results; they listed what the protocol did not decide. Those points were then decided as general rules from the coders' notes, by the time the second coding started and before

any result was counted, and applied to every case alike. A hash-chained work log, written by the coding tool's hooks and not by the AI sessions, recorded the time and hash of each file as it changed: the protocol at 12:28 UTC on 19 November 2026, the later decisions at 14:58 (added to the protocol at 15:27), and the brief that started the count at 15:53.

3 What we found

Measure	Claims	Share	95% interval
At least one counted check within 45 days	39 of 151	25.8%	19.5% to 33.3%
At least one counted check within 90 days	47 of 151	31.1%	24.3% to 38.9%
The proof itself examined within 45 days (reading, independent formalisation or exposition)	26 of 151	17.2%	12.0% to 24.0%
The proof itself examined within 90 days	31 of 151	20.5%	14.9% to 27.7%

For 55 of the 151 we found no record by anyone else of a kind the protocol notes down (bare praise, reactions and links are not noted).

Kinds. Within 45 days, 20 claims had a reading, 14 a statement review, 7 an exposition, 6 a re-run, 5 an independent formalisation and none a referee decision (a journal decision is not expected so soon; a claim can have more than one kind). Over 90 days there are 85 counted entries (one per person, place, day and kind), and 33 of them are statement reviews. Checking that a formal statement says what the problem asks is not the same as reading the proof.

Places. Within 45 days, 26 claims were checked in their own comments, 11 in Formal Conjectures pull requests, 6 in the problem's thread, 3 in the problem's notes and expositions and 2 in the erdosproblems repository. Later checks come more from GitHub, where the site's freeze of 6 October did not apply: the Formal Conjectures figure rises from 11 to 19 claims by day 90.

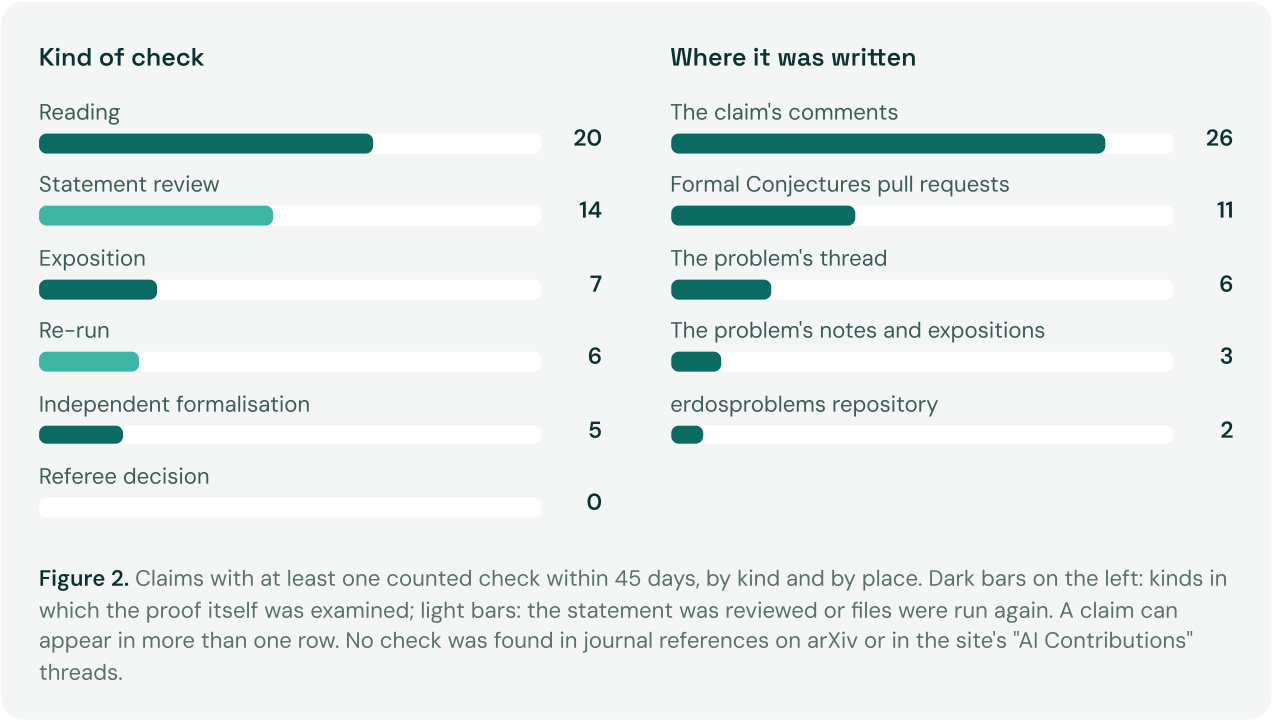


Figure 2. Claims with at least one counted check within 45 days, by kind and by place. Dark bars on the left: kinds in which the proof itself was examined; light bars: the statement was reviewed or files were run again. A claim can appear in more than one row. No check was found in journal references on arXiv or in the site's "AI Contributions" threads.

Speed. Where a check was found, it usually came fast. Of the 151 claims, 27 had their first counted check by day 7, 33 by day 14, 34 by day 30, 39 by day 45 and 47 by day 90. The median wait among those 47 is 6 days.

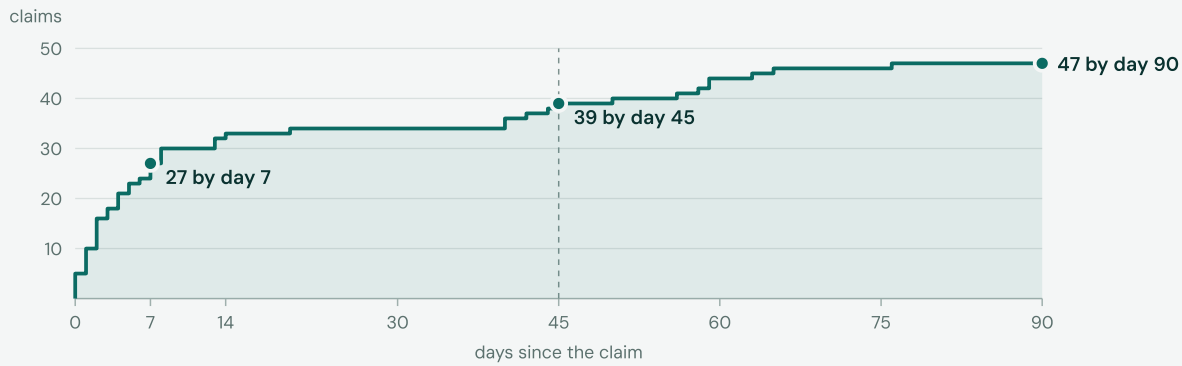


Figure 3. Claims that had received their first counted check, by days since the claim, out of 151. Most checks that were found came within the first week.

Who checked. Thirty-two different people wrote the counted checks of the first 90 days. The three most active account for 23 of the 73 pairs of a checker and a claim they checked (31.5%); without them the main number falls from 39 to 29. The site's owner checked 7 claims. Most checkers' profiles are empty, so this data cannot say who they are or whether they have an institution.

Checks made with AI. Twenty of the 85 counted entries (23.5%) say they were made with the help of AI: 7 of 8 independent formalisations, 7 of 28 readings, 5 of 33 statement reviews. For 7 of the 39 claims, every counted check within 45 days was of this kind.

The site's own mark. Eight claims carry the site's note that the proof was accepted as correct. We did not count the mark, because the page does not say who decided, when, or on what basis. Seven of the eight also have a counted check within 45 days.

4 How much the number moves

Seventeen readings of the rule that could move the main number were marked before the count, and the count was repeated with each taken the other way. The table shows the eleven that change it; six change nothing. Three later rules that could also move it have no other reading, and single judgements are not covered: two similar borderline cases were coded differently, and read alike the number would be 38 or 40.

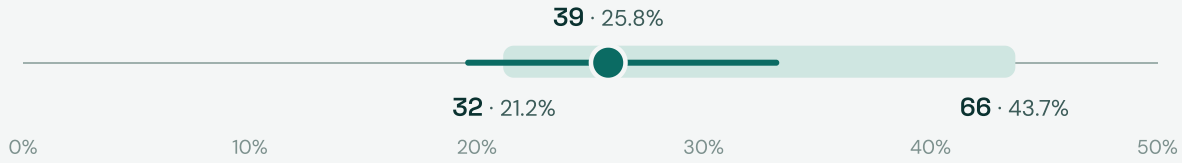


Figure 4. The main number under other readings of the rule, as a share of the 151 claims. Dot: the main measure. Line: its 95% interval. Band: from every reading that lowers it taken together to every reading that raises it taken together.

The other reading	Claims with a check	Change
An approval that does not say what was read counts	54	+15
A model's verdict passed on by someone else counts as a reading	47	+8
Compiling files without writing down the result counts as a re-run	45	+6
A scope objection that points to a specific part counts as a reading	42	+3
A formalisation by someone with a claim of their own on the same problem counts	40	+1
A check of someone else's formalisation of the claim does not count, and a check of formal files of unknown origin does	40	+1
A mismatch that lies in the repository's statement does not count	38	-1
The clock always starts at submission	38	-1
Day 45 is outside the window	38	-1
A check of a file that covers only part of the result does not count	38	-1
A formalisation that a third person made with AI does not count	37	-2

With every reading that lowers the number taken together it is 32 of 151 (21.2%); with every reading that raises it, 66 of 151 (43.7%). That span is wider than the 95% interval of the main measure. The largest single change comes from approvals that do not say what was read (+15), the next from relayed model verdicts (+8). There are many such approvals, and a reader cannot tell from them what was checked.

5 Two comparisons fixed in advance

We compared claims that link a formal proof with claims that do not (76 and 75), and claimants whose document or profile names an institution with those whose does not (34 and 65; for 38 more the class is unknown, and 14 link to a company's results page or are organisations).

Comparison	Measure	First group	Second group	Difference	95% interval
Formal proof linked / not	Any counted check	21 of 76 (27.6%)	18 of 75 (24.0%)	+3.6 points	-13.7 to +21.6
	Proof itself examined	10 of 76 (13.2%)	16 of 75 (21.3%)	-8.2 points	-22.2 to +7.7
Institution named / not	Any counted check	8 of 34 (23.5%)	14 of 65 (21.5%)	+2.0 points	-17.2 to +27.8
	Proof itself examined	4 of 34 (11.8%)	9 of 65 (13.8%)	-2.1 points	-18.2 to +17.7

Every interval contains zero, so we report no difference. The intervals are wide and do not rule out large differences either. These are comparisons of what was observed and show no cause. The intervals are from resampling the submitting accounts, because one account often submits many

claims; their ends move by 1 to 3 points with the random draw.

6 How far the coding can be trusted

All 154 cases were coded by AI sessions working from the written protocol. No person coded a case. That is a main weakness of this count.

A session on a different model coded 31 of the cases again. It was told not to open the first coding; the work log shows no step that did, but it does not keep the text of 52 of that session's 135 steps. The two agree on the main measure in 30 of 31 cases (Cohen's kappa 0.92); in 22 of the 31 neither found a check, so kappa rests on few cases. In the one case where they differ, the second coder did not record an exposition that stands on the problem's page; the first coder had missed it too at first and found it on a second pass. Both coders read the same protocol, so a shared misreading would not show up as disagreement.

A further session, on the same model as the first coders, then recomputed every figure from the data with its own code and re-read 16 of the 123 cases that only one coder had seen, eight with a counted check and eight without; it had read notes that mention some of them. In all 16 the main measure came out the same. Sixteen cases can only rule out a large error rate. A missed record can only make the true number higher than 39, not lower; a wrongly counted one would make it lower, and none was found in the 16.

7 Limits

- One site, one summer, mathematics only. Nothing here speaks for science as a whole.
- Only public records in the places listed. Email, social media, MathOverflow, discussions in the claimants' own repositories and journals' own sites were not looked at.
- The site froze on 6 October 2026, which cuts the site's part of the 90-day window short.
- Deleted claims are not in the archive, and we do not know which kind they were.
- The protocol was written after a 30-case pilot, by a session that knew how some of its rules played out in those cases.
- The later decisions were written after the cases had been coded, though before anything was counted. Their other readings are in section 4 too.
- The work log does not keep the text of commands passed inline (260 steps of the six sessions before the count), so it cannot show on its own that none of them added up results; that also rests on the sessions' instructions.
- The interval of the main measure treats claims as independent, although a few accounts submit many claims and a few people check many.
- Interest: Papex Labs builds tools for recording how research is made and checked, and lists AI-assisted claims. The answer to this question matters to it. Papex's own list was kept outside the main measure for that reason. The work log used here is Papex's own recorder, not yet released and here in its first trial on real work.

8 What we take from it

In this place, we found a recorded check by someone else within 45 days for about a quarter of the proof claims that name an AI system, and an examination of the proof itself for about a sixth. Where we found a check it usually came quickly, and a small group did much of the checking.

The second result is how soft the word "checked" is. The largest single swing in whether a claim counts as checked comes from records that say "looks right" without saying what was looked at. A check that names what was read, formalised or run can be seen and weighed by the next reader. An approval cannot.

Data and method

The protocol (in Turkish), the coding of every case, the second coding, the scripts and their outputs are deposited with this report at https://doi.org/10.5281/zenodo.24378081. People appear in the data only as numbers assigned at random; the key is not released. This keeps names out of the files but does not make people anonymous: each claim's number leads to its public page. The pages that were read can be found again from the claim numbers by the rule in the protocol; copies of them are not redistributed.

How this was made

The author set the question, approved the plan of the count and its change of scope, and decides what is published. AI sessions (Claude Code with Claude Opus 5.5 for the protocol, the first coding, applying the later decisions to every case, the analysis and the checks; Claude Sonnet 5.5 as second coder; a Claude session outside this folder as study manager, which wrote the briefs and the later decisions and drafted this text) wrote the protocol, fetched and read the pages, coded the cases and computed the figures. A separate Claude Opus 5.5 session checked this text against the report and corrected it. Where this note says "we", the reading, coding and counting were done by these sessions. The author read the result and is responsible for it. AI is not an author.