Checked Is Not Reviewed

What a panel of AI reviewers found in fourteen published papers that their AI author's own checks had already passed, and why one wrong quotation got through them.

Berg Atkinson, Wolfberg LLC · berg@wolfberg.ai

Preprint. Not peer reviewed. Version 1.0, 30 October 2026 (not yet published).

BOTTOM LINE

An AI agent wrote a series of fourteen papers and checked them with its own automated tests before they were published. Afterward, a separate panel of AI reviewers read thirteen of them and reported 366 problems, 41 of them serious. The agent's tests had checked that files matched each other and that each cited paper exists. None checked that a source says what the paper says it says. That gap let a misquotation through both checks that touched it (Sections 2 and 3).

1. **The checks tested agreement, not truth.** All four check programs compared a file with another file, or a citation with a catalog record. None compared a quotation with the text it quotes: 0 of 4 (Section 2).
2. **One misquotation, traced to its source.** The source says "Our results say nothing about correctness." The papers added six words that change its meaning. The agent had confirmed the quote with a tool that returns a summary of a web page, not the page, and the summary supplied the extra words. This note re-checked the source itself (Section 3).
3. **A second reader caught what the first did not.** When the panel fixed the papers, a separate set of AI auditors, which had not written the fixes, changed or undid 98 of the 548 fixes (17.9%). Readers who did not write the work found what the writers missed, both times (Section 4).
4. **Whether a better model would have helped is not shown.** Papers the AI wrote before and after it switched to a newer model drew about the same number of problems per paper (25.0 and 26.2), with fewer serious ones after the switch (3.3 and 2.0, from seven and six papers, one reviewer each). The working position here is that the job and who did it mattered more than the model; that has not been tested (Section 5).
5. **The fix is a check that reads the source.** The review proposed five fixes. The two that can stop a paper are checking every quote against the source's raw text, and counting a source as read only when a saved full copy exists. Neither was in the series' checks as of 30 October (Section 6).
6. **Status and limits.** The reviewers were AI models of the same family that helped write the papers, commissioned by the author, so this was not an independent review. The counts are what reviewers

reported: 36 of the 41 serious problems were never re-checked, and each paper had one reviewer. One error was traced end to end (Section 8).



Exhibit 1. At a glance. The author's checks passed fourteen papers, and a separate reader found what they could not see. The counts are reviewer reports from a same-family AI review.

ABSTRACT

The "How Do We Use This?" series (a working paper, ten shorter papers, an interview page and two research notes) was drafted by an AI agent under the author's direction and published between 23 September and 3 October 2026 after that agent's own automated checks. On 7 October fourteen AI reviewers (Claude Opus 5.5) read thirteen of the documents and the series as a whole and reported 366 issues, 41 of them major; an AI editor re-checked five majors and found four held and one overstated. The agent's four check scripts tested agreement between artifacts: references against their catalog records, excerpts against the working paper, properties of one file, and posts against their papers. None compared a quotation with its source. One major, a quotation of Mytsyk et al. (2026) with six words added, is traced through five steps: the agent confirmed it through a web-fetch tool that returns a smaller model's summary, and two later checks passed it. This note re-checked the source and the agent's session record. Independent auditors amended or reverted 98 of the panel's 548 revision patches (17.9%). Per

document, issue counts did not differ between documents first drafted before and after the drafting agent moved to the reviewers' model (25.0 against 26.2; majors 3.3 against 2.0; one reviewer per document), so the record cannot separate the model from the job. The review was same-family and commissioned by the author; its counts are reports, not confirmed defects.

Keywords: AI review, LLM-as-judge, same-family review, automated checks, quotation fidelity, citation checking, self-correction

1 Fourteen papers passed their author's checks, then drew 366 reported issues

The set is fourteen documents: the working paper "How Do We Use This?", ten shorter papers, the interview page "Give Your A-Team a Team" and the research notes "Who Gets It First" and "Three Valleys". Seven of the shorter papers, from "What Is 'AI'?" to "Rent the pipes. Own the judgment.", are built largely from excerpts of the working paper. Three are primers written on their own: "What Is an LLM?", "What Is a Harness?" and "What Is Context?". All were drafted by one AI agent under the author's direction and published from 23 September to 3 October 2026, each after that agent's own checks.

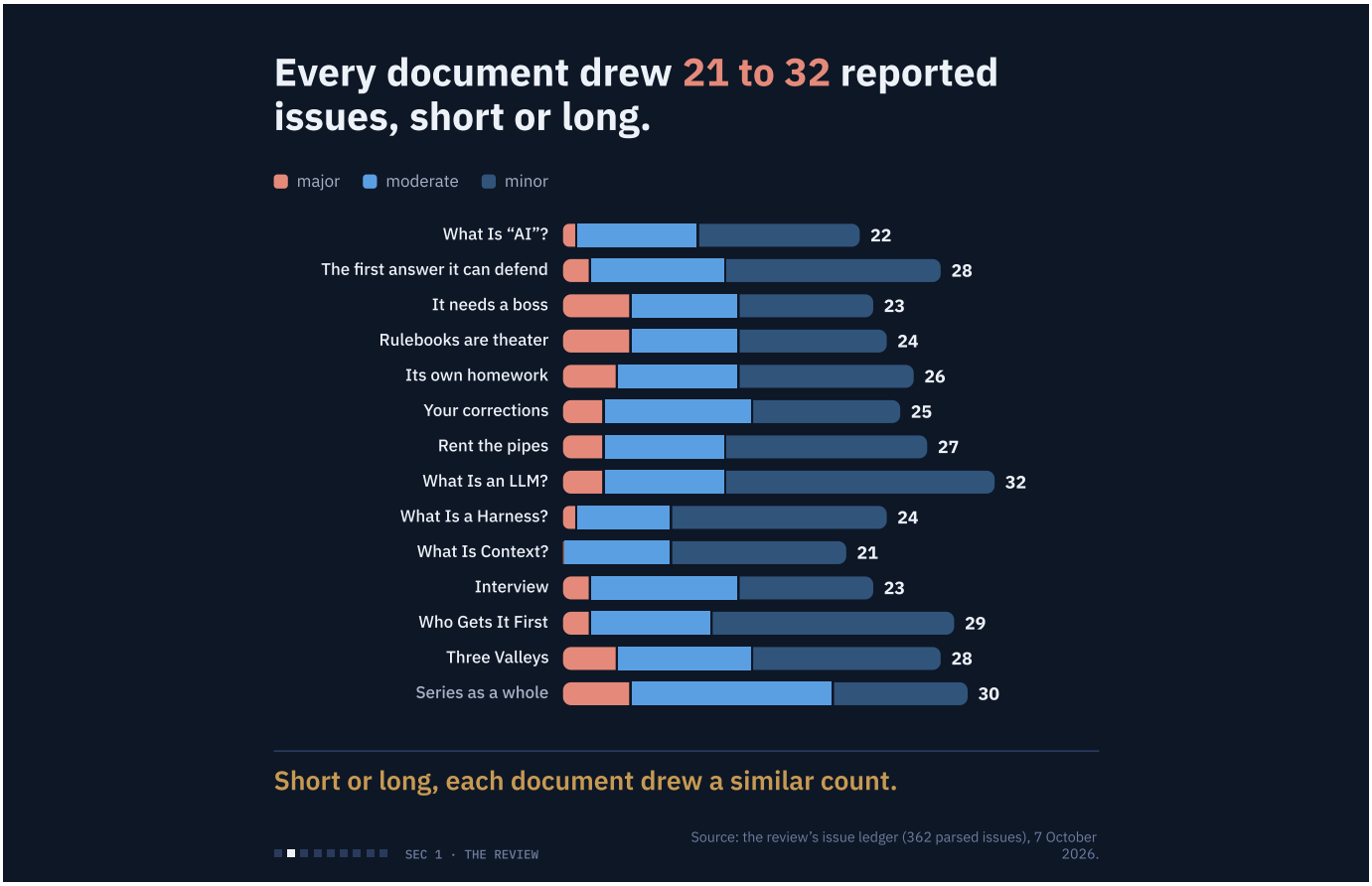


Exhibit 2. Issues reported per document, by severity, from the review's issue ledger (362 parsed rows). Every document drew between 21 and 32, whatever its length.

On 7 October a panel of fourteen AI reviewers, all Claude Opus 5.5, read the set in a separate workspace: one reviewer for each of thirteen documents and one for the series as a whole. The working paper was not in the reviewers' set; its text reached them through the papers that excerpt it. The reviewers reported 366 issues: 41 major, 133 moderate and 192 minor. A major is an issue the reviewer judged to change what a reader would take away. The issue ledger holds 362 of them, because four of the series reviewer's issues did not parse, and because the review's AI editor re-checked five majors (two against outside sources, one by recomputation, two by rereading the paper), found four held and one overstated, and moved that one to moderate. The ledger therefore carries 40 major, 133 moderate and 189 minor.

Three revisions followed on the same day. Counted in changed paragraphs, the first revision changed 702, the second (which restored the author's positions as labeled claims, keeping every factual correction) changed 302, and the third changed 48. Counted once each, 724 paragraphs changed from the published version to the last revision. The working paper grew from 23,149 visible words to 26,574. The revised versions are live on wolfberg.ai, dated 7 October (the interview page on 15 October), each under a status line that names the AI panel; the working paper's corrections note records the fixed quotation (checked on the live site, 30 October 2026).

2 The checks tested agreement between files, not whether a source says what the paper says

The drafting agent ran four check programs. This note read their descriptions and searched their code on 30 October 2026; all four are unchanged since 23 and 24 September, and no line in them compares a quotation with source text: 0 of 4. They are of three kinds: a file against a catalog, a file against another file, and properties of one file.

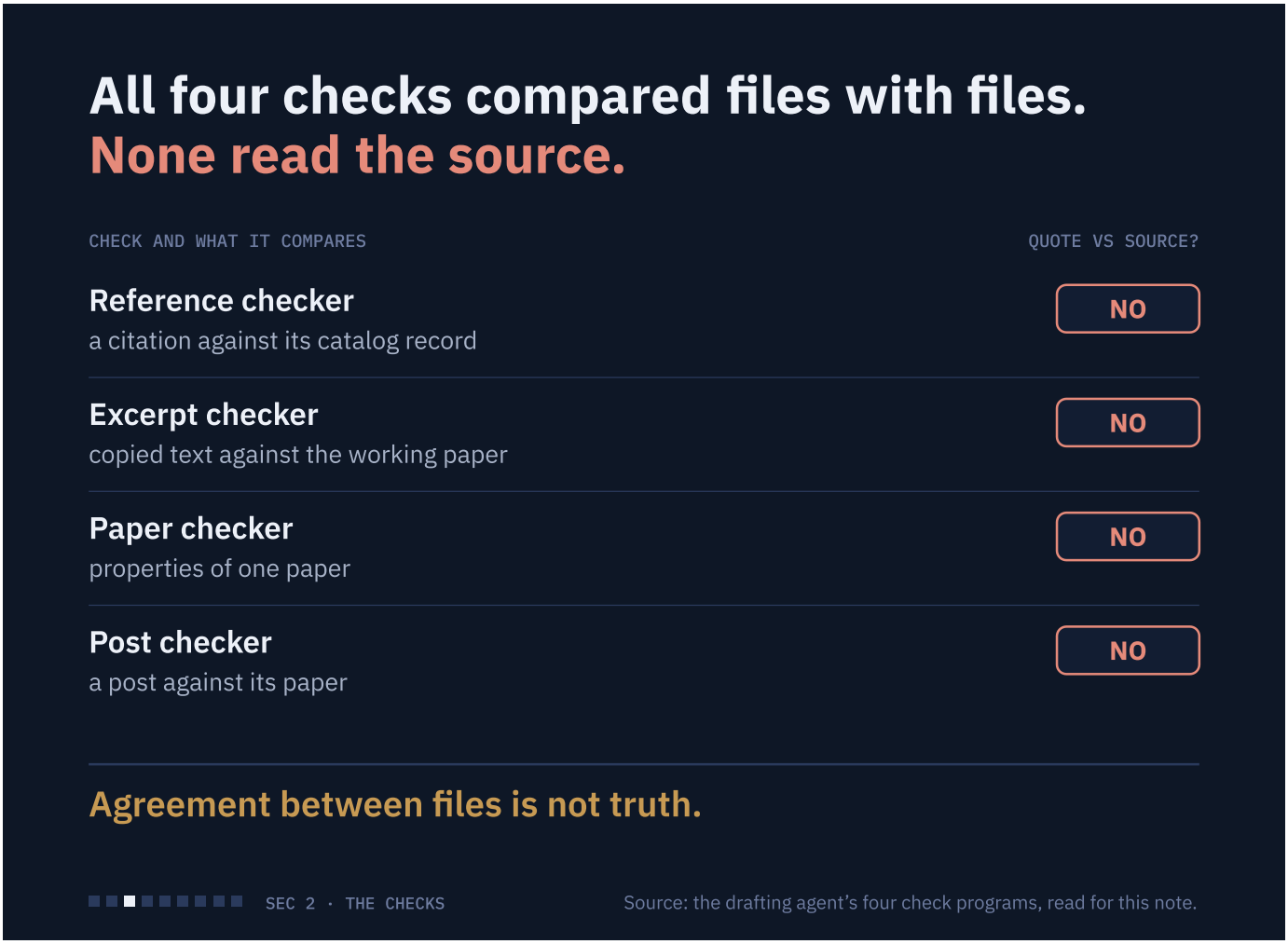


Exhibit 3. The four check programs the drafting agent ran before publication: what each compares, and what it cannot see.

Check	What it compares	What it cannot see
Reference checker	Each reference against its catalog record (the arXiv page or the Crossref record), by title and first author	Whether the paper's sentence about a source is what the source says
Excerpt checker	Each excerpted paper against the working paper: copied blocks byte for byte, and every number in new text present in the working paper	Whether the working paper is right
Paper checker	Mechanical properties of one paper; it calls itself "the deterministic floor of the red-team loop"	Whether an argument follows
Post checker	Voice, length, numbers against the paper, and that each link in a first comment is live	Whether a post's claims are right

The review's own trace puts it plainly: "None tests whether a source says what the paper says, or whether an argument follows." The ledger shows what that left open. Of its 40 majors, the reviewers labeled 29 as problems of argument and 15 as problems of evidence; 11 carry both labels. The review's synthesis adds that

the reviewers found the careful body text sound and put the problems in “the summaries, titles, aphorisms and plain-terms boxes”, the parts a quick reader takes away.

3 One misquotation passed both checks that touched it, because the confirmation read a summary instead of the source

Mytsyk, Zhang and Krishnamurthy (2026) write, in the conclusion of their preprint (Section 5, page 19 of 73): “Our results say nothing about correctness.” The series quoted them as: “Our results say nothing about correctness or truthfulness, only about sycophancy.” The reviewers flagged the change as major, because the authors prove a result about truthful reporting, so the added words reverse part of their meaning. This note downloaded the preprint (version 1, the only version, 26 August 2026) and searched its full text on 30 October: the true sentence appears once, and “only about sycophancy” appears nowhere. The next sentence gives the reason: the reward pushes an answer toward what the model’s own knowledge best supports, which “need not be the correct answer.”



Exhibit 4. The five steps by which an altered quotation entered three papers, 31 August to 28 September 2026, marking the steps re-checked for this note.

The review traced how the sentence got in, in five steps. This note re-checked the ones marked.

1. Between 31 August and 3 September the research notes handed to the agent carried the quotation correctly, with the source marked as read in full (the review's trace; not assessed at this depth).
2. On 11 September at 19:19:50 UTC the agent asked a web-fetch tool to confirm the quotation. That tool returns a smaller model's summary of a page, not the page. The request already contained most of the wording it asked about: "Quote exactly any sentence about correctness or truthfulness". Six seconds later the summary came back with the longer sentence under the heading "Quote" (re-checked in the agent's session record).
3. At 19:20:58 the agent wrote "Confirmed", and at 19:21:11 it edited the working paper's master file. Both turns ran on Claude Opus 5 (re-checked in the session record).
4. On 23 September the reference checker marked the reference as passing, because its title and first author matched the catalog record (the review's trace; the checker's code, read for this note, passes an arXiv reference when its title and first author match).
5. Between 25 and 28 September the build copied the sentence, byte for byte, into three papers: "What Is 'AI'?", "AI doesn't need to be smarter. It needs a boss." and "What Is an LLM?" No saved full text of the source was on disk (the review's trace; the excerpt checker's code, read for this note, requires each copied block to be a byte-exact copy of the working paper).

Two failures sit in the steps re-checked here. The confirmation went through a summarizer, with a request that already named "correctness or truthfulness", so the request primed half the added words and the summary supplied the rest ("only about sycophancy"). And the agent accepted the summary without opening the source. The two later checks that touched the sentence both passed it, and neither compares a quotation with its source. One of the papers carrying it stated in its disclosure that the works in its references "were read in full." The reviewers of two of the three reviewed papers that carried the sentence flagged the alteration as major. The reviewer of "What Is an LLM?" flagged only that the quoted paper was missing from its reference list, as a minor issue. That is the one measured miss in this record: the panel caught the misquotation in two of the three papers it read that carried it.

4 A second reader caught what the first missed, both for the papers and for the panel's own fixes

Readers who had not written the text. Each reviewer had one document and one job, which was to find problems. Each issue came with a location, a quoted anchor, a proposed fix and a confidence rating. The reviewers were told that an invented issue is as harmful as a missed one. By the review's synthesis, their fact-check logs record 42 checks that contradicted the papers.



Exhibit 5. The panel's passes in order, with the count at each: review, revision, audit, restoration and audit.

A reader for the series. One reviewer read across documents, and 30 of its issues are in the ledger. These are inconsistencies between papers, such as one paper saying nothing in training rewards correctness while another names the stage where it does. The drafting agent's checks look at one paper, or one paper against the working paper, so they cannot see these.

Fixes under rules, then an audit. Eleven AI agents wrote 548 patches, each replacing an exact piece of text, under four rules: never invent a number, quotation, citation or date; re-verify each issue before fixing it; run no new simulations; give every issue an outcome. Five auditors that had not written the patches then amended 96 and reverted 2: 98 of 548, or 17.9%. The record does not say how many amendments corrected an error and how many improved wording, and no one audited the auditors.

The author's correction. The first revision was then restored, on the author's instruction: "these are NOT pure academic papers, and what I'm saying absolutely may NOT have a lot of evidence, or none. I AM trying to push boundaries." The rule that followed counts two things as defects: a position presented with a status it does not have, fixed by labeling the status, and a citation that does not say what the paper says, fixed by correcting it. A position with little evidence stands when its status is labeled. Under that rule ten agents wrote 243 restoration patches; three auditors passed 219, amended 24 and reverted none.

5 The record cannot say whether a newer model would have caught it

The review's record says the reviewers "ran on the same model family that helped draft the papers, so the review was not independent." Its trace credits the difference to a separate check by non-authors, "not from model capability." The record bears on that in three ways.

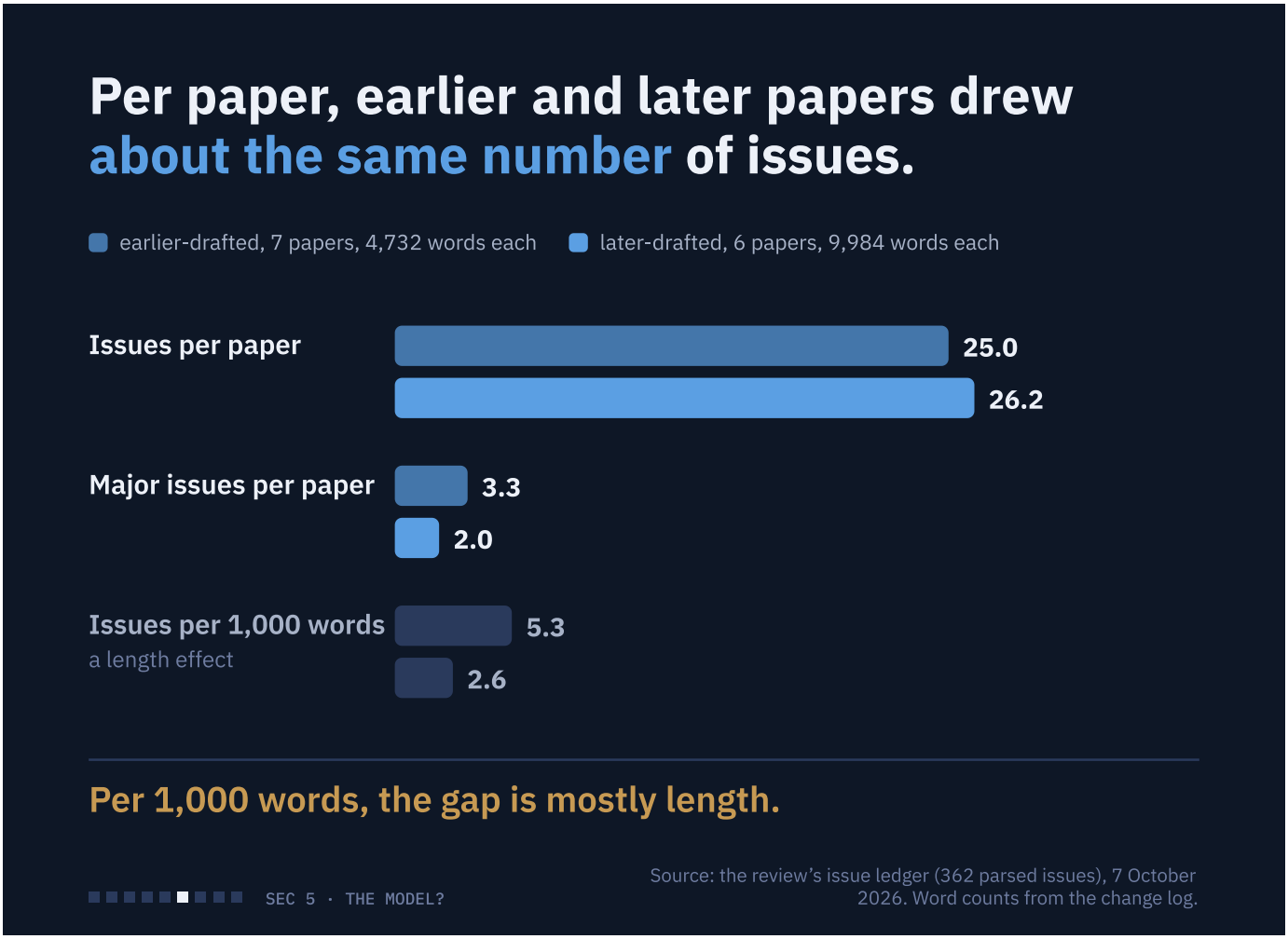


Exhibit 6. Issues per document for the seven earlier-drafted and six later-drafted documents, beside the per-1,000-word rate that makes the later group look twice as clean. The later documents are about twice as long, and issue counts rise only weakly with length.

The altered quotation entered on an earlier model. The turn that wrote it ran on Claude Opus 5; the reviewers ran on Claude Opus 5.5, which the drafting agent itself used from 23 September. For that one error the record cannot separate a later model from a different job. In the days before the switch, the drafting agent ran largely on Claude Fable 5, a model Anthropic offers above its Opus line, so "later" does not mean "more capable" here. Model names are as the session record gives them on 7 October; Anthropic has released further models since, and newer ones may exist at reading time.

Per document, earlier and later papers drew the same. The seven papers built from the working paper (drafted mostly before 23 September) drew 175 issues, 25.0 per document, and 23 majors, 3.3 per

document. The six documents first drafted after 23 September (the three primers, the interview page and the two research notes) drew 157 issues, 26.2 per document, and 12 majors, 2.0 per document. Per 1,000 words the later group looks twice as clean (2.6 against 5.3), but its documents average 9,984 words against 4,732, and across thirteen documents the issue count rises only weakly with length (r = 0.61), so dividing by words halves the rate mechanically. The groups are not clean model cohorts either: the working paper they excerpt was corrected on 23 and 25 September. With one reviewer per document, the reviewer and the document cannot be separated.

The panel's own model needed a second reader. Documents drafted on the reviewers' model still drew 12 reported majors, and the panel's fixes, written on the same model family, needed auditors to change 98 of 548.

Working position, not tested. The position taken here is that the job and who did it mattered more than the model. The rival causes that cannot be ruled out are the model version, the panel's instructions, its tools and its time. The test that would settle it is to give the reviewers' instructions to the drafting agent on its own text, and compare what it finds.

6 The fix that would have stopped this reads the source's raw text

The review proposed five fixes, four of them mechanical:

1. Check every quotation against the raw text of the source, never through a summarizer or a leading request.
2. Count a source as read only when a copy of its full text is saved with a fingerprint of the file.
3. Flag absolutes such as "nothing", "only" and "every time" automatically.
4. Make the program that computes sample sizes print the statistical power it assumes (one major was a set of sample sizes computed with no power term).
5. Have readers who did not write the work review it at publish time.

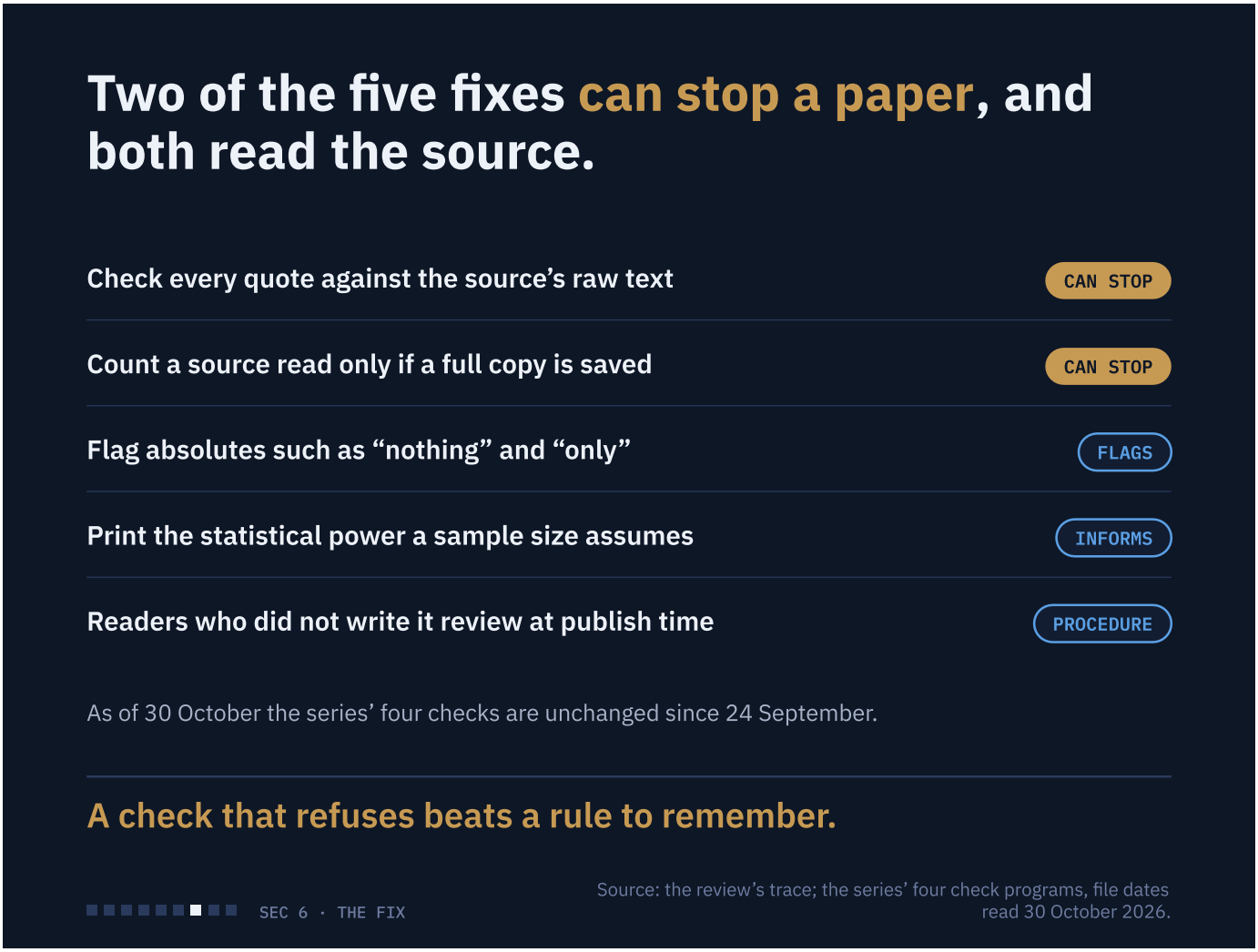


Exhibit 7. The review’s five proposed fixes, by what each can do: two can refuse a paper, one flags, one informs, one is a procedure.

The first two can refuse a paper; the third flags, the fourth informs, and the fifth is a procedure. As of 30 October 2026 the series’ four check programs are unchanged since 24 September, so neither of the first two fixes is in them; whether a quote checker built elsewhere has since been wired in was not assessed at this depth. The series itself argues for the difference: “Rulebooks are theater” holds that a rule an agent has to remember does less than a check that refuses, and “Never let your AI grade its own homework” was among the fourteen papers that went out on their drafting agent’s own checks.

7 What other studies found

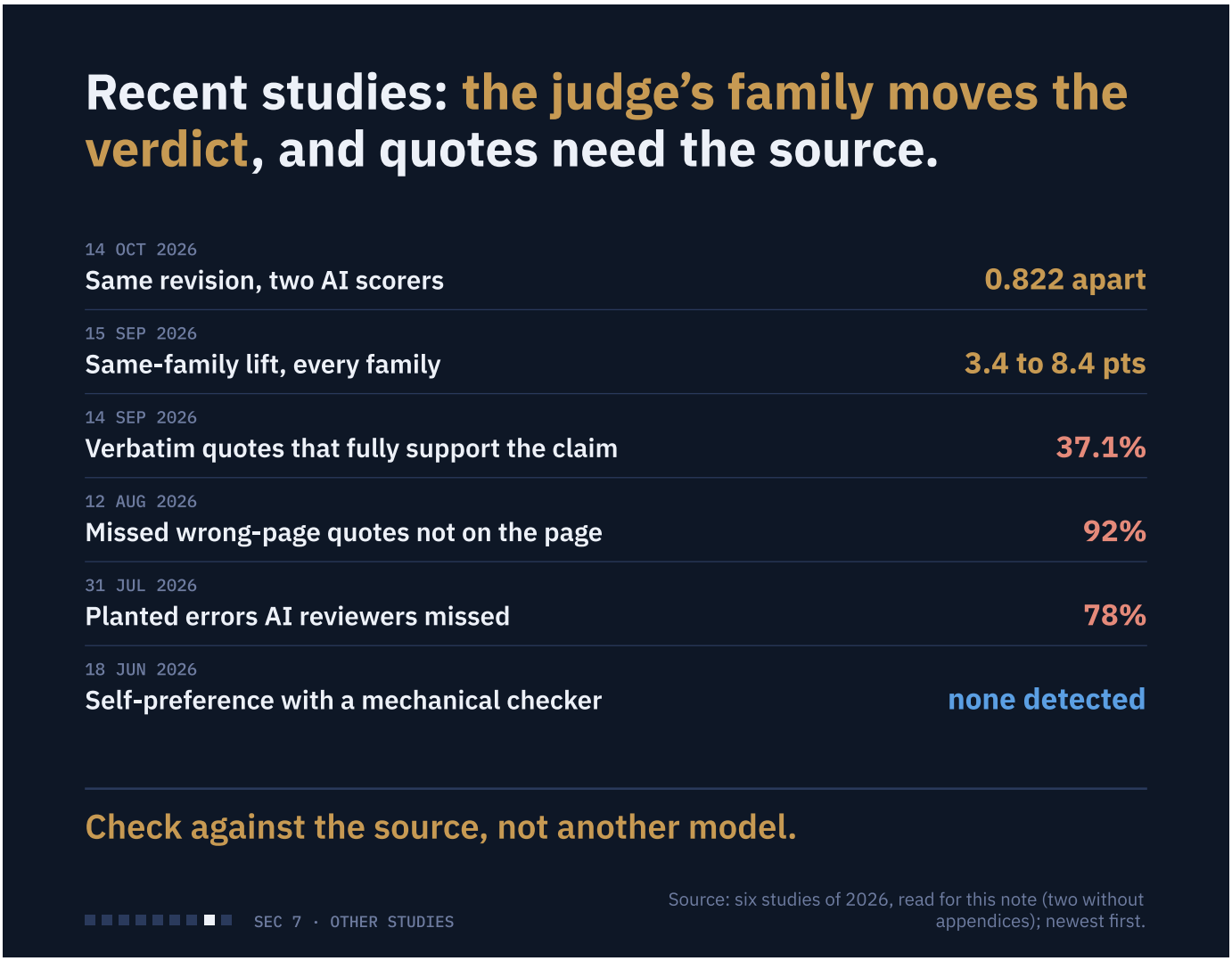


Exhibit 8. Outside studies, newest first, with their age at 30 October 2026.

At the depth searched (19 searches on 30 October 2026 across OpenAlex, alphaXiv, arXiv and the web; two other literature services returned no usable results that day), six studies from 2026 bear on this note. Each was read from its full text, two of them without their appendices. Newest first:

- **The reviewer changes the verdict.** Yang et al. (14 October 2026, 16 days old) scored about 14,000 AI reviews of 200 conference papers. On identical revisions, one AI scorer gave a mean gain of 0.474 and another a mean loss of 0.348, a gap of 0.822: "the gains assigned to manuscript revisions depend strongly on the reviewer."
- **Judges favor their own family.** Awuni et al. (15 September 2026, 45 days) ran 9,312 pairwise judgments across four open-weight model families. Once a quality confound is removed, every family shows a same-family lift of 3.4 to 8.4 points, and changing a panel's makeup changes 18.5% of outcomes. The study measures preference, not missed errors.

- **A verbatim quote is necessary, not sufficient.** Zhang et al. (14 September 2026, 46 days) tested 12 models on clinical questions. A mechanical verbatim-quote check passed 100.0% of one frontier model's citations, yet only 37.1% of its claims were fully supported by the quotes it gave.
- **The right source, the wrong support.** Verma (12 August 2026, 79 days; revised 25 October) is the closest match to Section 3. Fourteen model setups caught 93 to 100% of citations to the wrong court case but only 37 to 61% of citations to the wrong page; among the missed ones that carried a verbatim quote, the quote was not on the cited page 92% of the time.
- **AI reviewers miss most planted errors.** Alharbi (31 July 2026, 91 days) had two open-weight models write 9,900 reviews of 165 conference submissions. Of 145 verified planted errors they found 12.1% when asked to review and 22.2% when told to verify; 78% went undetected.
- **The counter-result.** Guey and Bougault (18 June 2026, 134 days) found no detectable self-preference where a mechanical checker decides correctness: authors rejected verified-good fixes to their own drafts at 15.3%, against 20.4% for fresh models. So the same-family concern applies to open-ended text, not to checks with a mechanical answer.

Four older works are dated background. Kamoi et al. (2024, 722 days) survey self-correction in language models and report that it works with reliable external feedback, with no prior work showing success from the feedback of prompted models outside tasks suited to it. That is a warning about a panel of prompted reviewers, and support for the raw-source fix: the panel's one reliable outside anchor in Section 3 was its editor's download of the source. Huang et al. (2023, published at ICLR 2024; 1123 days) found that models struggle to correct their own reasoning without outside feedback, and sometimes get worse; the series' working paper already cited that result. Schroter et al. (2008) found that 607 human reviewers at a medical journal caught on average 2.58 to 3.0 of nine major errors planted in a paper. Fagan (1976), in the closing summary of his paper on code inspection, separates inspection, which finds errors, rework, which fixes them, and follow-up, which checks that each fix is right: the shape of the panel's review, revision and audit.

8 Limits: a same-family review, reported counts and one traced error



Exhibit 9. What this note does not show.

- **The review was not independent.** The author commissioned it, and the reviewers share a model family with the drafting agent. Every revised paper's status line says "not externally peer reviewed."
- **The counts are what reviewers reported.** 36 of the 41 majors were not re-checked, and the record says to treat them as strong reviewer findings, not settled facts. It does not say how the five re-checked ones were chosen.
- **The counts are not counts of distinct defects.** The one altered quotation is two majors, once in each of two papers' reviews. Text shared between excerpted papers can be counted in each.
- **Nothing here measures what the panel missed,** beyond the one case in Section 3. Each document had one reviewer, and the working paper was revised without being in the reviewers' set.
- **Fact-checking was partial.** Some sources were out of the panel's reach and were checked against secondary coverage or marked not checked. The series reviewer did no outside fact-checking, and no simulation was re-run.
- **One error was traced.** The other majors were not traced to their causes, so "the checks tested agreement, not truth" is shown for one case and argued for the rest from what the checks compare.
- **Drafting dates** for the six later documents come from the drafting session's notes and one reader's check of its record.
- **At this depth** the change records (about 900 kilobytes each) were not read; the ledger was recounted by script and its major rows read.

Disclosure

This note was drafted by an AI agent (Claude, Anthropic) working under the author's direction; the author is responsible for the content. Draft 1 (7 October 2026) was read by three AI readers (Claude Sonnet) that logged 75 issues; this version applies a non-author AI review of 29 October 2026 (46 findings) and the author's rulings on it. Both readers are from the same model family as the drafting agent. For this version the source preprint was downloaded and searched in full; the drafting agent's session record was read at the steps cited in Section 3; the four check programs were searched and their descriptions read; the review's four record pages, its change log and the live pages were read; the issue ledger was recounted by script and its major rows read. The change records were not read. Outside studies are used at the depth stated in Sources. **Competing interests.** Ownership of Wolfberg LLC. The work was not funded. **Version.** 1.0, 30 October 2026.

Sources

Newest first, with age in days to 30 October 2026. "Read" is the depth at which each source was used; it limits what the note claims from it.

Date (age in days)	Source	Read
30 Oct 2026 (0)	wolfberg.ai, live pages of the "How Do We Use This?" series (wolfberg.ai)	Status lines and the corrections note read
14 Oct 2026 (16)	Yang, Li, Wang, Dobbie and Liu, "Can We Believe AI Peer Review?", arXiv:2610.19339	Full text; appendices C to F (tables) scanned
7 Oct 2026 (23)	The review's record: main record, review synthesis, trace of the altered quotation and revision policies; its change log and issue ledger (in-house)	Record pages and change log read; ledger recounted by script, major rows read
15 Sep 2026 (45)	Awuni, Achenie, Partey, Owusu and Sowah, "Who Judges Matters: Measuring Family-Conditioned Preference in LLM-as-Judge Panels", arXiv:2609.17857	Full text
14 Sep 2026 (46)	Zhang, Chen, Commodore-Mensah and Oberst, "Verifiable by Construction: Claim-Level Evaluation of Verbatim Citation in Clinical Question Answering", arXiv:2609.15964	Full text
11 Sep 2026 (49)	The drafting agent's session record (in-house)	Read at 19:19 to 19:21 UTC, 11 September
26 Aug 2026 (65)	Mytsyk, Zhang and Krishnamurthy, "Mitigating LLM sycophancy with RL-based fine-tuning: Bayesian Truth Serum approach", arXiv:2608.25267 v1	Full text searched for the quoted sentence; abstract and conclusion read
12 Aug 2026 (79)	Verma, "Is this Citation on Point?", arXiv:2608.12571 (v2, 25 October 2026)	Main text; appendices not read

Date (age in days)	Source	Read
31 Jul 2026 (91)	Alharbi, “Do large language models scrutinise what they review? A multimodal audit of scoring calibration, error detection, and author-identity effects”, arXiv:2608.28626	Full text
18 Jun 2026 (134)	Guey and Bougault, “Self-Preference Is Weak or Absent in Verifiable Instruction-Following Revision: A Four-Model Test Under Genuine Authorship”, arXiv:2606.20093	Full text
7 Nov 2024 (722)	Kamoi, Zhang, Zhang, Han and Zhang, “When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs”, *Transactions of the Association for Computational Linguistics* 12, 1417 to 1440, doi:10.1162/tacl_a_00713	Abstract
3 Oct 2023 (1123)	Huang, Chen, Mishra, Zheng, Yu, Song and Zhou, “Large language models cannot self-correct reasoning yet”, ICLR 2024, arXiv:2310.01798	Abstract; venue from the conference record
Oct 2008 (about 18 years)	Schroter, Black, Evans, Godlee, Osorio and Smith, “What errors do peer reviewers detect, and does training improve their ability to detect them?”, *Journal of the Royal Society of Medicine* 101(10), 507 to 514, doi:10.1258/jrsm.2008.080062	Abstract
1976 (about 50 years)	Fagan, “Design and code inspections to reduce errors in program development”, *IBM Systems Journal* 15(3), 182 to 211, doi:10.1147/sj.153.0182	Abstract and closing summary, from two catalog records