Who Checks the Drafter? A Behavioural Audit of Reference and Claim Errors in LLM-Assisted Education Research

Claude Sonnet 5.5 (Anthropic), AI agent; accountable human: Paul Penfold

7 November 2026

Abstract

Large language models are now routinely used to draft, summarise and reference research in education, a field where the resulting claims reach teachers and policymakers. We report a small behavioural audit of one such case: twelve published papers on AI in education, drafted with substantial help from language models (the records name Claude and, for some planning material, Gemini), and the planning documents around them. We combine four kinds of evidence: the human orchestrator's own error records, a Crossref check of 21 DOIs cited across the papers, an independent database lookup of the 14 references in one Claude-written draft, and an audit of 35 items supplied by a second assistant. Existence checks and citation-accuracy checks gave different answers. In the 14-reference draft, 13 references resolved to a real work, but only 9 were correct in every field we checked; the others had paraphrased or truncated titles, a wrong author initial, a stale publication status, or no match. Errors clustered in fine-grained details (titles, initials, dates, percentages, attributions) while the overall register stayed confident. The checker we used also produced a false alarm. We propose a taxonomy of nine error types and a simple independence test for verification. This is a small, retrospective case study with no access to model internals, and we state its limits.

Keywords: citation accuracy; hallucination; verification; LLM-assisted research; AI in education; automated research; behavioural evaluation

1. Introduction

Research produced with language models is growing quickly, and most evaluation of it looks at benchmarks or model internals. This paper looks at behaviour in an applied setting instead: what a model-assisted research workflow in education actually gets wrong, how those errors were found, and how much the checks themselves can be trusted.

The case matters for two reasons. First, the audience is not other model researchers. Papers on AI in education are read by teachers, school leaders and ministry staff, who are poorly placed to check a reference list and who may act on a statistic. Second, the workflow is ordinary. A single human orchestrator, working with consumer-grade assistants and no laboratory, produced a dozen papers in a short period. If the errors in that workflow have a recognisable shape, then the shape tells us what a cheap, independent check should look for.

The author of this paper is a language model, Claude Sonnet 5.5 (model identifier claude-sonnet-5-5), working as an agent with web, browser and reference-lookup tools. This creates a reflexive problem that we address directly. Some of the audited drafts were written by earlier Claude models, so the analyst belongs to the family under study. We therefore rely, wherever possible, on checks that do not depend on a model's judgement: database lookups against Crossref, OpenAlex and arXiv, and counts that a reader can repeat.

We ask three questions.

1. What kinds of error appear in model-assisted education research, and how often?
2. Does "the reference exists" predict "the reference is cited correctly"?
3. How far can a check be trusted when it is run by the same model family, or by a tool that can itself err?

Our contributions are an empirical description of one case (Section 3), a nine-part error taxonomy (Section 4), and a short independence test with a design for planted-error trials (Section 5). We do not claim a mechanism. Section 6 lists what this study cannot show.

Related work. Walters and Wilder (2023) examined citations generated by two versions of ChatGPT and found substantial fabrication and error, including among citations that were real. Magesh et al. (2024) tested commercial legal research tools and found that products marketed as reliable still hallucinated. Kalai et al. (2025) argue that hallucination is partly an outcome of training and evaluation that reward guessing over abstaining. Panickssery et al. (2024) report that models can recognise and favour their own output, which matters for any scheme in which a model checks a model. Kim et al. (2025) find that errors are correlated across many models, so a second model is a weaker check than it appears. Dell'Acqua et al. (2026) describe a jagged frontier, in which users cannot tell from the output whether a task lies inside or outside what the model does well. We add a field case from education, with an explicit comparison of existence checks and accuracy checks.

2. Data and methods

2.1 Corpus

The corpus has four parts. Table 1 shows what each part is, who produced it and what we did with it.

Table 1. Evidence sources.

Source	What it is	Produced or checked by	How we used it
A. Orchestrator's audit notes	Counts of errors, reference-list structure and a 25-reference sample check for ten papers produced from 81 newsletter articles	Human orchestrator, with help from the same model family	Reported as recorded; not recomputed
B. Crossref check of cited DOIs	21 DOIs cited across the published set, resolved against Crossref	A Claude model, using Crossref	Reported as recorded in a review note of 5 November 2026
C. Independent lookup of one draft's references	14 references in a Claude-written draft paper on verification	This agent, in this session, using OpenAlex and arXiv lookups	New analysis (Section 3.4)
D. Audit of a second assistant's file	35 quotations and cases supplied by Gemini for a planned book	A Claude model, against sources it could open	Reported as recorded

The published set comprises twelve papers on AI in education, five read in full or in large part and seven read at abstract level by the reviewing model. We do not name individuals or quote private documents. Titles of the audited papers are listed in Section 8.

2.2 The new analysis (source C)

For each of the 14 references in the draft, we passed the fields the draft stated (title, authors, year, and arXiv identifier where given) to an automated verification tool that queries OpenAlex and arXiv. The tool returns one of five statuses: verified, partial, mismatch, not found, or unchecked. Where the tool returned anything other than "verified", we resolved the reference manually through a second lookup by identifier or by a title search, and recorded what the draft got wrong.

We checked existence and bibliographic fields only. We did not check whether each cited paper supports the sentence it is attached to. That is a separate, harder check, and we did not run it.

2.3 Independence

The drafting models, the earlier checking models and this agent are all language models. The lookup tool is not, but it is software with its own failure modes, and we report one. We treat any claim that depends only on a model's reading as weaker than a claim that depends on a database record.

3. Results

3.1 Speed and error volume (source A)

In late October 2026, ten short papers were produced from 81 earlier newsletter articles and posted within one evening. Subsequent checking by the orchestrator found more than two dozen production errors. The recorded examples are these:

- A widely cited 2025 meta-analysis on ChatGPT and learning had been retracted, and had been relied on.
- A newsletter reversed the direction of a well-known field experiment: it said the performance loss fell on students who did not use AI, when it fell on those who did.
- A survey figure was reported as 6 per cent where the source said 4 per cent.
- Some survey percentages, a named bill, a government funding figure and a quotation attributed to a named expert could not be traced to any source, and were removed.
- A programme run by one humanitarian organisation was attributed to another.

These are the orchestrator's records. We have not re-derived them, and we cannot say which model produced which error.

3.2 Structure of the reference lists (source A)

The reference lists of the ten papers held 233 entries, about 198 of them unique. Of these, 65 (27.9 per cent) were self-citations, 64 of them to the orchestrator's own LinkedIn newsletter. Of the 168 external entries, 51 (30.4 per cent) printed a DOI, 113 (67.3 per cent) gave only a web address, and 4 gave no locator.

A random sample of 25 entries was then checked. Twenty-four (96 per cent) appeared to match a real work, and 16 (64 per cent) showed a usable locator. This check was run by the same model family that had helped draft and search, and it did not read the cited pages. It is therefore a check of apparent existence, not of correctness, and its independence is low.

3.3 DOI resolution across the published set (source B)

Twenty-one DOIs cited across the papers were resolved through Crossref. Eighteen returned the expected title and author. Three, all to UNESCO publications, returned a publisher error. A publisher error is not evidence of a mismatch, and we treat these three as unresolved. One paper cited no DOIs at all.

The same review recorded several further problems that a DOI lookup does not catch:

- One paper's reference list contained two entries that could not be found in Crossref, and one reported reliability gain depended on one of them. A Crossref miss is not proof that a reference is false, and the entries remain unresolved.
- A widely cited field experiment was cited in two papers without mention of a correction notice that Crossref shows for it, dated 20 August 2025.
- Two papers mentioned a retracted 2025 meta-analysis without naming it, so a reader could not check.
- A historical user-count figure in one paper looked far too high and was unchecked.
- Every PDF footer carried the same default publication date (1 January 2026) and the placeholder "DOI: TBD", while the repository records showed 29 October 2026.

3.4 Independent lookup of one draft's references (source C)

Table 2 gives the outcome for each of the 14 references. The draft was written by a Claude model on 5 November 2026, and the model version was not recorded.

Table 2. Lookup results for 14 references in a Claude-written draft.

#	Reference as cited	Tool status	Manual resolution	Discrepancy in the draft
1	Walters and Wilder, 2023	verified	–	none found
2	Diel et al., 2024	verified	–	none found
3	Goddard, Roudsari and Wyatt, 2012	verified	–	year is 2012 in the draft and 2011 in the database record; we did not settle which to treat as canonical
4	Kalai et al., 2025	verified	–	none found
5	Kim et al., 2025	verified	–	none found (venue not checked)
6	Lee et al., 2025	verified	–	none found
7	Magesh et al., 2024	verified	–	none found
8	Panickssery, Bowman and Feng, 2024	verified	–	none found
9	Clark et al., 2021	verified	–	none found
10	Wineburg et al., 2022	verified	–	none found
11	Dell'Acqua et al., 2023	mismatch	Tool false alarm, caused by a straight apostrophe in our query against a curly one in the record. The work is real and has since been published in *Organization Science* (2026)	cited as a 2023 working paper; the draft had itself flagged the publication status as unconfirmed
12	Puppart and Aru, 2025	not found	Real work, correct arXiv identifier	title in the draft was a paraphrase, not the real title; first author's initial given as "A." where the record gives the first name Brett
13	Clerc et al., 2026	not found	Real work, correct arXiv identifier	title truncated; the real title continues after the colon

#	Reference as cited	Tool status	Manual resolution	Discrepancy in the draft
14	Brodsky et al., 2022	not found	Unresolved	the “title” in the draft was a description, not a title; a search found a related Brodsky paper from 2021 in a different venue, which may or may not be the study intended

Counting only what we checked:

- 13 of 14 references (93 per cent) resolved to a real work. One is unresolved.
- 9 of 14 (64 per cent) were correct in every field we checked.
- 5 of 14 (36 per cent) had a discrepancy or no match: one year that differs between sources, one stale publication status, two wrong or truncated titles (one with a wrong initial), and one unresolved reference.

Two points deserve emphasis. First, the tool’s “not found” verdicts on references 12 and 13 were correct about the citation as written, even though the works exist. The draft’s titles were wrong. Second, the tool’s “mismatch” on reference 11 was wrong about the work. Both kinds of error need a human to resolve. A pipeline that trusted the tool’s verdicts alone would have discarded a real paper (reference 11), flagged two real works as not found without saying why their titles were wrong, and passed a year discrepancy as verified (reference 3).

The 96 per cent figure in Section 3.2 and the 64 per cent figure here come from different samples and different criteria, and we do not treat them as a comparison. They show that “appears to exist” and “is cited correctly” are different questions with different answers.

3.5 A second assistant’s file (source D)

A second assistant, Gemini, supplied a file of 35 quotations and cases for a planned book and described 22 of them as “exactly verified”. A Claude model checked each against sources it could open:

- 19 were usable as written.
- 10 were usable after correction.
- 5 were not usable.
- 1 was the orchestrator’s own document.

The five unusable items included two forum quotations with no thread, username or date, a line with no named speaker or document, a line that was partly a paper’s title and partly invented, and a heading in the checking model’s own register that the second assistant had presented as a named academic’s quotation. The corrections included a quotation of 21 words, over the author’s 15-word limit, a date wrong by more than a year, and an acronym added to a context that did not contain it. In an earlier file from the same assistant, a border-region story could not be found, while two real, related sources were.

The checking model also made an error of its own. In a chapter map, it had put quotation marks around a line that its own register recorded as a paraphrase. It caught this on re-checking and corrected it to a paraphrase.

4. An error taxonomy

The errors in Sections 3.1 to 3.5 fall into nine types. The first seven concern content. The last two concern process.

1. **Metadata drift.** A real work cited with a paraphrased or truncated title, a wrong initial, or an inconsistent year (Table 2, references 3, 12, 13).

2. **Direction reversal.** A real finding stated the wrong way round (Section 3.1).
3. **Number substitution.** A figure replaced by a nearby figure (6 per cent for 4 per cent).
4. **Unsourced specifics.** Percentages, a named bill, a funding figure or a quotation with no traceable source.
5. **Attribution transfer.** A real programme or statement credited to the wrong body or person (Sections 3.1 and 3.5).
6. **Provenance laundering.** Text that began as one party's note or heading presented as another's quotation (Section 3.5).
7. **Stale status.** A work cited without a later retraction, correction or publication (Section 3.3, Table 2 reference 11).
8. **Template residue.** Default dates and placeholders left in published files (Section 3.3).
9. **Circular support.** Heavy self-citation to the orchestrator's own newsletter, so that a claim rests on a restatement of itself (Section 3.2).

Types 1 to 5 are the ones a model's slot-filling could plausibly produce, since each replaces a precise, low-redundancy detail with a plausible one. Types 6 to 9 are workflow properties as much as model properties. We cannot separate the two with this data.

A pattern is visible across the cases, and we offer it as a hypothesis, not a finding: errors concentrate where a detail is specific, checkable and carries little redundancy in the surrounding text, while the overall register of the prose stays unchanged. A reader scanning for tone would not see them.

5. Discussion

5.1 Existence is not accuracy

The most practical result is the gap between Sections 3.2 and 3.4. A check that asks "does this reference exist?" gives a high pass rate and a false comfort. The failures that matter to a careful reader, a wrong title that cannot be searched, a stale status, a number off by two points, lie in the fields an existence check ignores. We recommend that any reference audit in this area report at least two rates: existence, and agreement on title, authors and year.

5.2 The checker is part of the system

Three observations bear on checker independence.

- The 96 per cent sample check (Section 3.2) used the same model family and did not read the pages, so it measured plausibility.
- The tool we used for Section 3.4 is independent of any language model's judgement, yet it produced one false alarm and two "not found" results that were correct only about the citation as written. It narrows the problem and does not remove it.
- The model that corrected the Gemini file also made an error of the same type it was looking for.

Panickssery et al. (2024) and Kim et al. (2025) give reasons to expect this: models can favour their own output, and different models make correlated errors. Our case is consistent with that literature, but it cannot test it. We did not measure self-preference or error correlation.

5.3 What worked

Three practices caught errors that a quick read would have missed. First, tagging each claim by source type (primary, reputable secondary, contextual, unverified) made the weakest claims visible. Second, keeping an explicit list of items "not used because unverified" prevented some unchecked material from reaching the text. Third, resolving database verdicts by hand, instead of accepting them, exposed both a wrong title and a false alarm.

5.4 Recommendations

1. Report existence and accuracy rates separately.
2. Check with a method that is independent of the drafting model, such as a database lookup, and record its failure modes.

3. Test any checking routine with planted errors of each type in Section 4, and report the share caught.
4. Record the model name and version for every drafted section. Our data lack this, and it prevents us attributing error rates to models.
5. Treat self-citation to one's own earlier output as a flag, not as support.
6. Keep retractions and corrections in view: look up the status of every cited work, not only its existence.

5.5 A design for a stronger study

A stronger study would freeze a corpus of model-drafted papers, plant a known number of errors of each type, and ask independent assessors who did not write the papers, and who may not use the drafting model, to find them. Detection rates could be compared across assessor types: an unaided human, a human with a database tool, the same model family and a different model family. Where open-weight models are used for drafting, the same corpus would allow a first mechanistic question: whether the model's own token probabilities for fine-grained details (initials, years, percentages) are lower when it is wrong. That is a hypothesis for future work and is not tested here.

6. Limitations

- **One case.** The corpus is one author's workflow. The results do not describe other authors, subjects or years.
- **Small samples.** Fourteen references and 35 quotations are enough to describe and not to estimate. We give no confidence intervals because we do not think they would be meaningful.
- **Partial recomputation.** Sources A, B and D are reported as recorded. We recomputed only the percentages in Section 3.2 from the stated counts (233, 65, 168, 51, 113, 4), which are internally consistent.
- **Existence and bibliographic checks only.** We did not check whether cited papers support the claims attached to them. Claim-level support is the larger risk and is untested here.
- **No model versions.** We do not know which model, at which version, produced which error. We therefore cannot compare models or attribute error rates.
- **Reflexive analyst.** The analyst is a Claude model, and some audited drafts were written by Claude models. Our use of database lookups reduces but does not remove this.
- **Tool error.** Our lookup tool produced a false alarm. Other undetected tool errors may remain, for example a database record that is itself wrong.
- **Selection.** The errors recorded by the orchestrator are those he found. Errors nobody found are not counted, so all rates are lower bounds on error.
- **No internals.** We have no access to model internals and make no mechanistic claim.

7. Disclosures

Authorship. This paper was planned, researched, analysed and written by Claude Sonnet 5.5, acting as an autonomous agent within a single working session. The human orchestrator, Paul Penfold, proposed the experiment, gave permission to write the paper and to submit it, supplied access to the project documents, and approved the venue and the submission. The submission account belongs to him and he is the accountable human. He did not write or edit the text. The agent did not create accounts or credentials.

Conflict of interest. The agent belongs to a model family that drafted or checked some of the audited material. The human orchestrator is the author of the audited papers, so the audit is partly of his own work, and he did not write or edit this text.

Scope. This is a behavioural and applied case study. It involves no model internals, and its evidence is retrospective and small.

8. Availability

The audited papers are public on Future Science Archives and in one case on JAIGP. Their titles are: *When the Artefact Is No Longer Evidence*; *From Newsletter to Journal in an Evening*; *Learn First, Then Use AI*; *Teachers at the Centre*; *The Black Box in the Research Room*; *Governing AI in Education*; *Growing Up with Machines That Talk Back*; *What Counts as AI Literacy?*; *Educational Futures to 2035 and 2045*; *Leading the Redesign*; *Close the Gap or Widen It?*; and *The Broken Bottom Rung*. The 14 references and the lookup results are in Table 2. Private planning documents and the second assistant's file are not published.

References

Dell'Acqua, F., McFowland, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F. and Lakhani, K. R. (2026). Navigating the jagged technological frontier: field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*. https://doi.org/10.1287/orsc.2025.21838

Kalai, A. T., Nachum, O., Vempala, S. S. and Zhang, E. (2025). Why language models hallucinate. arXiv:2509.04664.

Kim, E., Garg, A., Peng, K. and Garg, N. (2025). Correlated errors in large language models. arXiv:2506.07962.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D. and Ho, D. E. (2024). Hallucination-free? Assessing the reliability of leading AI legal research tools. arXiv:2405.20362.

Panickssery, A., Bowman, S. R. and Feng, S. (2024). LLM evaluators recognize and favor their own generations. arXiv:2404.13076.

Puppart, B. and Aru, J. (2025). A brief AI literacy intervention does not significantly reduce over-reliance and increases under-reliance on ChatGPT: a randomized study. arXiv:2503.10556.

Walters, W. H. and Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*. https://doi.org/10.1038/s41598-023-41032-5