

Peer Review

Review of: "An AI-Human Collaborative Review of AI-Human Collaborative Reviews"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

This is a single-author narrative review of how researchers and generative AI tools can co-produce scholarly text. The question is stated at the end of the **Introduction**: "How can AI and humans collaborate with each other for academic writing, what are the possible perils and pitfalls, and what are some recommended steps to mitigate such risks?" The **Materials and Methods** section describes a three-tool pipeline. Ai2 Paperfinder was given one prompt, reproduced verbatim in the text ("Identify the current status, benefits, drawbacks and recommended practices for using generalised pre-trained models for collaboration between humans and Artificial Intelligence"). ResearchRabbit was then used to expand that seed set and to locate full texts. Google NotebookLM was used to extract, against three stated questions, the authors, the main messages, and the conclusions of each retrieved PDF. The Abstract is explicit that "AI tools were not used for content authoring."

The **Results** are organised as five subsections: benefits of co-writing, best practices, whether AI can be a co-author, limitations and caveats (hallucination, bias, paper mills, red teaming), and remedies, the last of which summarises the PaperCard framework [18] and the 2023 STM white paper. The **Discussion** lands on three positions: human-AI collaboration produces better and faster scholarly output, AI cannot be an author, and authors must disclose AI use and keep it to editorial assistance.

Let me record what the paper does well, because some of it is unusual. The reflexive design is a genuine idea: a review of AI-assisted scholarship that discloses its own AI-assisted method and can therefore be audited as an instance of the thing it describes. The verbatim search prompt and the three extraction prompts are printed, which is more than most narrative reviews offer. The summaries of PaperCard [18]

and of the STM advisory read accurately against what those documents actually say, including the point about readers not uploading published manuscripts to public platforms, which is usually left out. The closing limitations paragraph is candid about the reliance on AI search and about the absence of structured selection criteria. My objections below are not that the author is unaware of the weaknesses. They are that the claims made in the Abstract and Discussion are stronger than the method can carry, and that the reference list undercuts the paper's own central methodological assertion.

2. Whether the conclusion holds on the data

2.1 The corpus is never counted. This is the defect that governs everything else. Nowhere in **Materials and Methods** or **Results** is there a number: not how many records Paperfinder returned, not how many ResearchRabbit added, not how many PDFs were uploaded to NotebookLM, not how many studies were finally included. There are no tables and no figures in the version rendered on Qeios. The only countable artifact is the reference list, which carries 18 numbered entries, and two of those ([4] Cole & Boutet on ResearchRabbit, [5] Tufino on NotebookLM) are tool citations rather than evidence about human-AI collaboration. So the evidential base is at most 16 items, and the Results discuss roughly a dozen. A reader cannot distinguish between a review that screened 300 papers and kept 16 and one that kept everything the first query returned.

2.2 "Consensus" rests on four sources, and the author disqualifies one of them himself. The first sentence of **Benefits of teaming up with AI to write papers** states that "in the corpus of studies reviewed, there was a consensus that when humans and AI collaborate, the overall impact is greater than the sum of the parts." Four studies follow: Veach & Abualkibash [6], Lin [7], AbdElsalam & Abdel-Momen [8], Nguyen et al. [9]. Of the first, the author writes, correctly, "this does not quite qualify as an AI-assisted paper being written; rather, here, AI is itself the sole author of a paper." That leaves three sources supporting the claim, one of which ([8]) is itself a commentary rather than a study. Consensus is the wrong word for agreement among three opinion pieces selected by a single query.

2.3 No study characteristics, no designs, no effect sizes. Nguyen et al. [9] is reported as showing that doctoral students in "iterative, highly interactive" workflows "tended to achieve better performance in their writing tasks" than students using a linear style. How many students, what design, what outcome measure, what magnitude: none of it appears. The same holds for Reza et al. [17], where the one quantitative detail given is the size of their own corpus (109 HCI papers, 2018 to 2024), not any result. The single outcome number in the whole review is from Kacena et al. [15]: "the AI-only approach resulted in

up to 70% of the cited references being inaccurate." Even that arrives without a denominator, without the number of articles generated per arm, and with "up to", which reports a maximum rather than a rate. I cannot check any of these figures against the text because the text does not contain them.

2.4 Critical appraisal is claimed but not shown. **Materials and Methods** states that "both peer-reviewed articles and preprints were considered as knowledge objects, and both types of publications were critically appraised before summarisation." No appraisal instrument is named, no appraisal results are given, and no study is described as having been downgraded or excluded on quality grounds. Since roughly a third of the reference list is preprints ([5], [10], [14], [17], [18] all carry arXiv DOIs), the appraisal step is exactly where a reader would want detail.

2.5 The hallucination-freedom claim is broader than the argument that supports it. **Materials and Methods** asserts that "the workflow described here for generating this review is free from the risk of producing fictitious or hallucinatory research publications," and gives three reasons: Paperfinder draws on Semantic Scholar, ResearchRabbit checks DOIs, and NotebookLM reads only uploaded files. Those three reasons bear on one question only, whether the cited papers exist. They say nothing about the second and larger risk, which is whether NotebookLM's extraction of "the main messages" and "the conclusion" of a PDF is faithful to that PDF. Retrieval-augmented systems misattribute claims inside real documents. The paper does state that "all uploaded materials were manually read, and data were abstracted and combined with insights generated from NotebookLM to verify the sources of information," which is the right control. But the result of that control is never reported. How many extractions did the manual read contradict? Zero? Three? That single number would have turned a demonstration into evidence, and it is the most valuable thing this study could have produced. Its absence is why the design remains an uncontrolled case report of one manuscript.

2.6 The reference list contradicts the weaker, checkable version of the claim. I went through all 18 entries. None appears fabricated, which supports the narrow point about Semantic Scholar indexing. But the metadata is visibly unchecked, and in this paper, that matters more than it would anywhere else.

- Reference [11] reads: "Lee, J. Y. (2023). Can an artificial intelligence chatbot be the author of a scholarly article? Null. <https://doi.org/10.3352/jeehp.2022.20.6>". The journal field is the literal string "Null", and the DOI encodes 2022 against a stated year of 2023. This is the source on which the entire **Can AI be treated as a co-author?** subsection rests.
- Reference [7], Lin, is dated 2023 while its DOI is 10.1038/s41551-024-01185-8, a 2024 registration.
- Reference [2], Zhang et al., is dated 2023 and cited to *Briefings in Bioinformatics* 25(1), a 2024 volume.

- Reference [3] is rendered in full capitals with a spelling error carried over from the source metadata: "IS ARTIFICIAL INTELLIGENCE HALLUCINATING?"

Four entries out of eighteen show defects consistent with automated metadata import that no one read afterwards. Free of fabrication is not the same as accurate, and a paper arguing that a tool chain protects citation integrity needs its own citations to be clean.

2.7 Stated scope and delivered scope disagree. The opening sentence of **Materials and Methods** says the purpose was a narrative review "for research on health and related sciences." The research question in the **Introduction** carries no such restriction, and neither does the corpus: the included work spans human-computer interaction [10], [17], higher education [9], agricultural science [8], bioinformatics [2], and publication ethics [12], [13]. Either the health framing should go, or the selection should be restricted and re-run.

2.8 The search is one query on one system. A single prompt to Paperfinder, expanded by one seed-based recommender, with no date limits, no stated search date, no inclusion or exclusion criteria, and no second database. The author concedes the point in the final paragraph ("future studies need to utilise both AI with prompts and search engines"), and I agree with the concession, but it does not neutralise the sensitivity problem: with one query and no screening log, there is no way to estimate what the review missed.

2.9 The seed prompt contains a terminology error that may have shaped retrieval. Both the Paperfinder prompt and the surrounding text use "generalised pre-trained" models. GPT expands to *generative* pre-trained transformer. The error repeats at least four times (twice in the **Introduction**, twice in **Materials and Methods**, once inside the quoted prompt itself). Elsewhere the paper writes "artificial generative pre-trained transformer-based applications," which is a third variant. In prose, this is a slip. Inside the literal query string that generated the entire corpus, it is a methodological fact, and the paper should either re-run the query with the correct term and report whether the hit set changed, or state that the wording was not varied.

2.10 The Abstract asserts an outcome that was never measured. The Abstract concludes that genAI tools combined with human authoring "can provide excellent exemplars of human-AI collaboration, particularly in improving the flow and quality of the output." Nothing in this paper measures flow or quality. No text was rated, no comparison arm exists, and the only exemplar available is the manuscript

itself, which is not assessed. This sentence should be rewritten as a report of what the included sources claim, not as a result of this review.

2.11 "Unanimous" on one citation. The subsection **Can AI be treated as a co-author?** opens: "The answer to this question is a unanimous 'no.'" One source is then cited, Lee [11]. The position is very likely right, and the ICMJE and COPE statements would support it, but unanimity across a single reference is not a finding.

2.12 The title overstates the corpus. "An AI-Human Collaborative Review of AI-Human Collaborative Reviews" promises a review of collaborative *reviews*. Of the included work, only Kacena et al. [15] concerns review articles. The rest is about academic writing in general, co-writing tools, and publication ethics. The second half of the title should say writing, not reviews, unless the corpus is narrowed to match.

3. Reproducibility

I want to give credit first, since the bar in this literature is low: the search prompt is printed word for word, and the three NotebookLM extraction prompts are printed as a numbered list. That is more prompt disclosure than most papers of this kind provide, and it is the part a future author could actually re-use.

No tool or model versions anywhere. NotebookLM is a front end over a Gemini model that has been swapped several times; the paper names neither the model nor the date of access. Paperfinder is described through a blog quotation, with no version and no date. ResearchRabbit likewise. Since none of the three services versions its outputs publicly, a run today will not reproduce a run from the paper's writing period, and the paper needs to say so rather than leave it implicit.

No access dates and no retrieval logs. The search date is not given. No export of the Paperfinder result set, no ResearchRabbit collection link, no list of the PDFs uploaded to NotebookLM. All three tools export; a CSV and a shared collection would have cost the author nothing and would have made the corpus auditable, which is the specific thing section 2.1 above asks for.

No decoding parameters, by a route the paper itself identifies. The **Limitations and Caveats** section is pointed about this: temperature is "usually hidden from the user unless the user uses an application programming interface," and consumer chat front ends leave the user with "no control" over pre-specified parameters, producing "a 'black box' and a level of opacity." That diagnosis is correct, and it applies to this manuscript in full. The paper cannot report temperature, top-p, context handling, or

retrieval settings for any of its three tools, and it does not acknowledge that its own method inherits the opacity it criticises two pages later. One sentence would fix the inconsistency.

Non-compliance with the paper's own recommended framework. The **Possible remedies and recommendations** section endorses PaperCard [18] and lists its component (4): authors should "specify the names, service providers, relevant dates, versions and specifications of the AI model(s) used, provide the prompts they used, and demonstrate their intellectual contributions." This manuscript supplies names, service providers, and prompts. It supplies no dates, no versions, and no specifications. The author anticipates the gap ("fulfilling all four requirements or even the suggestions of the framework are tall asks for many, if not most, authors"), but a paper recommending a standard is the one place where the standard should be met in full, and where meeting it would double as a worked example for readers.

Declarations not legible in this rendering. As rendered on Qeios, the Declarations block is a collapsed link, and its content did not appear in the page text I worked from, so I cannot confirm whether a data availability statement or a separate AI-use declaration exists. If there is none, one belongs there.

One technical description is wrong. **Limitations and Caveats** states that "mimicking humans is a function of adding perplexity parameters in the form of 'temperature' in the model." Temperature is not a perplexity parameter. It is a scaling factor applied to the logits before the softmax, which flattens or sharpens the next-token distribution. Perplexity is an evaluation measure of a language model over held-out text, not a setting an operator adjusts. Two distinct concepts have been merged into one phrase, and in a submission filed under Computer Science, that will be noticed.

A related dating problem sits in the same paragraph: "Generalised pre-trained models are built on different text corpora, including WebText and OpenWebText [14]." Those are GPT-2 era corpora. Presenting them as representative of the training data behind current systems is several model generations out of date, and the sentence should either be marked as historical or replaced.

4. What specifically to fix

1. **Materials and Methods, the sentence beginning "The texts were then summarised to generate themes on three aspects to explore".** Five aspects follow, labelled (i) through (v), and the Results are organised into five matching subsections. Change "three" to "five".
2. **Materials and Methods, new paragraph and a flow diagram or short table.** Report four counts: records returned by the Paperfinder query; records added by ResearchRabbit; full texts retrieved and uploaded to NotebookLM; studies included in the synthesis. Give the search date. Without these

numbers, the reader has no way to size the corpus behind the word "consensus" in the Benefits subsection, and the review cannot be assessed as a review.

3. Add a table of included studies, one row per source: first author, year, country or setting, design (empirical study, systematic review, commentary, editorial), sample where applicable, and the finding used in this paper. Roughly a dozen rows. It would immediately expose how much of the evidence base is commentary rather than data, which is a finding in itself and worth stating plainly.
4. **References [11], [7], [2] and [3].** Reference [11] carries the literal journal name "Null" and a year that disagrees with its DOI (10.3352/jeehp.2022.20.6 against a stated 2023). Reference [7] is dated 2023 against DOI suffix s41551-024-01185-8. Reference [2] is dated 2023 against *Briefings in Bioinformatics* 25(1). Reference [3] appears in full capitals with "ARTIFICIAL" misspelled. Correct all four by hand and say in the Methods that reference metadata was verified manually, since the paper's argument depends on it.
5. **Materials and Methods, the sentence "The workflow described here for generating this review is free from the risk of producing fictitious or hallucinatory research publications".** Narrow the claim to what the three reasons support, namely that cited publications are real and indexed, and add a separate sentence on extraction fidelity. Then report the number that the manual verification produced: of N extractions checked against the full text, M disagreed. Even N = 12, M = 1 is worth more than the assertion as it stands.
6. Replace "generalised pre-trained transformer" with "generative pre-trained transformer" throughout (Introduction, twice; Materials and Methods, twice). The instance inside the quoted Paperfinder prompt should stay as it was issued, with a note stating whether the query was re-run with the corrected term and whether the hit set changed.
7. **Limitations and Caveats, the sentence on temperature.** Rewrite: temperature rescales the output distribution over the next token, and perplexity is an evaluation metric, not a control. Update or date-stamp the WebText and OpenWebText sentence in the same paragraph.
8. **Discussion, first sentence.** "The infallible accuracy of a software tool" cannot stand in a paper whose **Limitations and Caveats** section documents fabricated references, algorithmic bias, and context-induced error. Say what is meant: mechanical consistency in spelling, grammar, and formatting.
9. **Discussion, end of the second paragraph.** The text stops mid-sentence at "In conclusion," and the next paragraph opens on limitations. Finish the sentence, or delete the fragment and give the paper a proper closing section.

10. **Materials and Methods, opening sentence.** Either drop "for research on health and related sciences" or restrict the corpus to health and re-run. At present, the stated scope and the included studies do not match.
11. **Results, "Can AI be treated as a co-author?".** Either drop "unanimous" or cite the policy statements that would justify it, which for this claim are the ICMJE recommendations, COPE's position on AI authorship, and the publisher policies referenced in the STM white paper already discussed in the remedies section.
12. **Abstract, final two sentences.** Attribute "improving the flow and quality of the output" to the reviewed sources, since this study measured neither.
13. **Title.** Change the second half to reflect the corpus, which is about collaborative writing rather than collaborative reviews.
14. Add a short Declarations entry meeting PaperCard component (4) as endorsed in the remedies section: tool names, providers, access dates, model versions where visible, the prompts (already present), and the author's own contribution. This is the cheapest fix on the list, and it makes the paper an example of its own recommendation.

5. Recommendation

Major revisions.

The topic deserves attention, and the reflexive design is worth keeping. The author writes clearly, reads the sources he cites, and is honest in the closing paragraph about the search being thin. Two things block acceptance in the present form. First, no reader can size or audit the corpus: there are no counts at any stage, no table of included studies, no appraisal results, and the word "consensus" is doing work that three commentaries cannot support. Second, the paper's strongest methodological claim, that the workflow is free from the risk of hallucinated publications, is stated more broadly than its own three-part argument covers, is never backed by the verification data the author says he collected, and is undercut by four defective entries in his own reference list, including a journal field that reads "Null" on the single source underpinning the co-authorship section.

None of this requires new research. The counts exist in the author's Paperfinder and ResearchRabbit sessions, the verification result exists in whatever he noted while reading the PDFs, and the reference corrections are an afternoon's work. Reporting them would convert a readable essay into a review with an

auditable method, and would let the manuscript do what its title promises: serve as a demonstration that survives being checked.

Declarations

Potential competing interests: No potential competing interests to declare.