CRITICAL / PRACTICE NOTES

Can you trust AI-generated summaries?

Verifying AI-assisted academic reading

NO. 05 / NOVEMBER 2026 ROY HANNEY READING TIME 6 MIN

The provocation

When we ask generative AI to summarise an academic paper, how do we know that the summary faithfully represents the knowledge contained within it? We might check that the references exist, ask for supporting quotations or instruct the AI to review its own work. But what exactly have we established through these activities?

A genuine reference can be attached to an invented finding. A reasonably accurate summary can nevertheless change the meaning of an argument by removing a qualification. Even a seemingly rigorous process of iterative self-correction may leave the original errors undetected.

Perhaps we need to reconsider what we are asking students to learn when they engage with AI-assisted academic reading. The educational opportunity may lie in making the relationship between knowledge claims and their supporting evidence an explicit object of enquiry.

The critical question

What would happen if we organised the reading process around the verification of individual knowledge claims, making their evidential foundations available for examination?

What the research suggests

Peters and Chin-Yee (2025) investigated nearly 5,000 AI-generated summaries of scientific research, examining changes in the scope of knowledge claims. These included the transformation of quantified findings into generic statements, the extension of findings beyond their original temporal context and the conversion of descriptive conclusions into recommendations for action (pp. 2–3, 15).

Of particular interest is their finding that a prompt explicitly requesting accuracy and comprehensiveness was associated with increased overgeneralisation compared with a simpler prompt (p. 10). An instruction to be accurate cannot itself be understood as a mechanism for establishing accuracy.

Linardon et al. (2026) examined 333 citation–claim relationships across twelve AI-generated literature reviews in mental health. Although all citations could be traced to identifiable publications, only 43.5% of the attributed claims were judged completely accurate. Almost one-fifth fell into the lowest accuracy category, which included unsupported or fabricated findings and claims contradicted by the cited source (PDF pp. 3–5).

The difficulty here is one of evidential alignment. Establishing that a publication exists tells us relatively little about whether it supports the proposition for which it has been cited.

Could we address these problems by asking AI to check and correct its own work? Kamoi et al. (2024), in their critical survey of self-correction research, found limited reliable evidence for intrinsic self-correction across general tasks under appropriately designed experimental conditions. They identify the generation of reliable feedback as a central difficulty. More promising conditions arise where verification is easier than initial generation, or where dependable external tools or evidence can assist the checking process (pp. 1423–1426).

The problem is therefore not simply whether an AI-generated summary is accurate, but how students might learn to examine the evidence on which its claims depend.

An approach to try

Invite students to investigate the fidelity of an AI-generated summary of an academic paper. Select a short, accessible reading appropriate to the discipline and provide everyone with the original full text.

01 ESTABLISH — Generate an initial reading. Ask students to use an AI tool to summarise the paper's argument, findings and limitations. Retain the original summary as a baseline output.

02 CHALLENGE — Investigate self-correction. Ask the AI to check its summary and correct errors. Retain the self-corrected version and a brief record of what changed. Did it identify errors or merely sound more convincing?

03 INTERROGATE — Construct an evidence record. Select five substantive claims from the summary. For each, evaluate whether the original paper actually warrants the claim. Record evidence with page references, or explicitly note when support is absent, ambiguous or spread across passages:

Generated claim	Source evidence and page(s)	Discrepancy or qualification	Revised claim
What does the AI claim?	What supports the claim, if anything? Note absent or ambiguous evidence.	Does the source warrant the claim? What is missing or altered?	How should the claim be represented?

Read each passage in context: it may report an earlier study, propose a hypothesis, offer an interpretation or acknowledge a limitation. Finding similar wording does not establish that the passage supports the generated claim.

Classify the discrepancy, where relevant. Has association become causation, a situated observation become a generalisation, or uncertainty disappeared? Is the evidence insufficient, contradictory or missing? Explain the judgement.

04 CALIBRATE — Correct and compare. Give the AI specific, evidence-grounded feedback and request revisions only to affected claims. Independently check the corrected claims against the paper, including for new errors. Retain the evidence-informed version.

Compare the initial, self-corrected and evidence-informed summaries, using the five-row verification record to explain differences.

Discuss which errors each approach exposed, which remained unresolved and what required students' own disciplinary judgement.

Where AI fits — and where it does not

AI can assist with extracting candidate claims, locating passages, identifying possible discrepancies and proposing revisions. These activities can make aspects of scholarly reading available for closer examination.

AI-located passages must still be read in context, and apparent discrepancies checked. A claim may rely on several passages or on an extended theoretical argument rather than a single sentence. Where the source does not warrant a claim, students should record that uncertainty rather than manufacture support.

The exercise is intended to develop students' capacity to make informed judgements about knowledge and evidence. AI becomes a participant in that enquiry, generating material which students can investigate, challenge and reconstruct.

Questions for adaptation

- How might this approach differ when reading a scientific paper, a philosophical argument or a piece of qualitative research?
- Could students use their evidence records as part of a literature review or dissertation preparation exercise?
- What might we learn about students' developing scholarly judgement by examining the discrepancies they identify and the reasons they give for correcting them?

Three sources to explore

Peters, U., & Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Royal Society Open Science, 12*(4), Article 241776. https://doi.org/10.1098/rsos.241776

Linardon, J., Messer, M., Anderson, C., Soliman, O. M., Liu, C., Firth, J., & Torous, J. (2026). Fluency without fidelity: Errors in citation-attributed claims in large language model-generated literature reviews in mental health. *Journal of Technology in Behavioral Science*. https://doi.org/10.1007/s41347-026-00630-6

Kamoi, R., Zhang, Y., Zhang, N., Han, J., & Zhang, R. (2024). When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics, 12*, 1417–1440. https://doi.org/10.1162/tacl_a_00713

Method & provenance

This note emerged from an AI-assisted critical enquiry into the reliability of AI-assisted academic reading, initially prompted by a practitioner discussion of source verification on The Slow AI Substack. The enquiry developed through structured analysis of three open-access research papers, comparison of their methodological contributions and iterative drafting. The full papers were consulted and relevant publication page numbers checked during the preceding analyses; PDF page numbers are used for Linardon et al. where stable journal pagination is unavailable.

The enquiry also informed the development of a researcher-directed annotated bibliography and literature synthesis framework. The pedagogical approach presented here is an interpretive contribution arising from that work. Its effectiveness as a teaching intervention has not yet been empirically evaluated.

Initial provocation: https://theslowai.substack.com/p/can-i-trust-chatgpt-check-source

ABOUT THIS SERIES

Short, evidence-informed provocations that move from critical enquiry to adaptable teaching practice. Each note draws on a small, accessible reading list and makes the role of AI-assisted scholarship transparent.