APPLING
International Applied Linguistics Research Conference

https://doi.org/10.29140/9781763711686-13

From Prompt to Citation: Auditing AI-Mediated Literature Synthesis in Applied Linguistics

Minhao Yuan
The University of Melbourne, Australia

Abstract

Generative artificial intelligence (AI) can assist literature synthesis, but fluent summaries and plausible citations may obscure how manuscript claims relate to original sources. This methodological paper proposes a prompt-to-citation audit framework for postgraduate research in applied linguistics. Adapting the qualitative audit-trail tradition and complementing publisher AI-use disclosure, the framework links prompt documentation, claim identification, source verification, and citation decisions. Verification authenticates references and their bibliographic details before checking whether sources support the proposed claims. Five provisional categories distinguish accurate, partial, overgeneralised, unsupported, and misleading claims. A hypothetical second language writing example shows how the audit can expose differences between perceptions, revision outcomes, and writing development. Recording model versions, session context, and archived outputs supports inspection of the process despite output variability. The paper also addresses workload through selective documentation of claims proposed for citation. The framework remains unvalidated; future evaluation should examine classification consistency, error detection, and the time required for verification.

Keywords: generative artificial intelligence; literature synthesis; audit trail; citation accuracy; applied linguistics

E-mail: yuanminhao202509@gmail.com

In Y. Wang, J. Matthews, & B. Zou (Eds.), *Proceedings of the 2nd International Applied Linguistics Research Conference, 2026* (pp. 99–104). Castledown.

Background

Generative artificial intelligence (AI) can support literature work for postgraduate researchers in applied linguistics. Tools such as ChatGPT can summarise articles, group studies, compare findings, suggest theoretical links, and help researchers move across large bodies of multilingual scholarship. Applied linguistics researchers often work across areas such as second language (L2) writing, feedback, multilingualism, and language policy. For multilingual postgraduate researchers in particular, AI tools may appear to reduce the burden of reading, organising, and writing across languages and academic conventions. In language education, Kohnke et al. (2023) discuss ChatGPT's affordances and responsible use; Barrot (2023) identifies response accuracy concerns in L2 writing.

However, AI-mediated literature synthesis also introduces a methodological risk: the evidential path between source texts and final citations may become blurred. A literature claim in a manuscript may appear to be supported by a citation, but the claim may actually have been produced through an AI-generated summary, an AI-proposed category, or the researcher's interpretation of an AI output. This paper is a methodological contribution rather than an empirical evaluation of AI tools. It asks how researchers can make visible the path by which a claim moves from prompt, to AI output, to source verification, to final citation.

Recent evidence reinforces the need for verification. Chelli et al. (2024) documented hallucinated references in biomedical systematic-review tasks. Clark et al. (2025) found substantial errors across evidence-synthesis tasks and limited robust evidence of time savings. These findings concern the evaluated models and tasks; they do not establish error rates for applied linguistics.

Literature reviewing requires explicit methodological decisions (Snyder, 2019) and iterative interpretation (Boell & Cecez-Kecmanovic, 2014). Building on Lincoln and Guba's (1985) use of audit trails for dependability and confirmability, the framework adapts this logic to preserving traceability between prompts, AI-generated claims, original sources and final citations. Elsevier (2026) requires disclosure of tools, purposes and human oversight, alongside output verification. Disclosure describes how AI was used; the proposed audit records how individual claims were checked, revised or removed. Its contribution is a practical adaptation linking each candidate claim to its prompt, output, source passage and final citation decision.

Prompt-to-Citation Audit Framework

The framework has four linked stages: prompt documentation, claim identification, source verification, and citation decision documentation (Table 1).

Table 1. Prompt-to-Citation Audit Framework

Stage	Researcher action	Audit question	Output
Prompt documentation	Record prompt, purpose, scope, source set, date, tool, model/version and available settings.	What was AI asked to do?	Prompt log
Claim identification	Mark each proposed summary, category, comparison, or theoretical link.	What claim may enter the manuscript?	Claim record
Source verification	Authenticate the reference; check and classify the claim against the original text.	Does the source exist and support the claim?	Verification note
Citation decision	Retain, revise, narrow, remove, or reattribute the claim and citations.	Which source supports the final claim?	Citation decision trail

The prompt log should retain the tool/provider, displayed model name and version, access date, exact prompts, conversation context, source files, raw outputs, and available generation or retrieval settings. Record hidden version details or settings as unavailable. Give prompts and claims identifiers linking them to verification notes. Archived outputs preserve the record for inspection even when sessions or model updates produce different text. The aim is procedural traceability.

Verification has two steps. First, match author, title, year, and identifier, where available, against an independent publisher or bibliographic record. Withhold unverified references; an inaccessible source is unresolved, not necessarily fabricated. Second, read the original text and record the supporting passage and location. A real reference alone does not establish support for a claim.

After source authentication and access, assess each candidate claim against its proposed sources. **Accurate** preserves the source's meaning, scope and qualifications. **Partial** contains supported and separable unsubstantiated components, without contradicting the source or extending its findings.

Overgeneralised extends a supported finding beyond its population, context, duration, outcome or causal warrant. **Unsupported** has no relevant supporting evidence in the checked sources. **Misleading** incorrectly reports what a source did or found, including its method, measured construct or direction of results. Split compound claims where possible. For overlapping labels, prioritise misreporting, then scope extension, then degree of support; record the passage and rationale. These provisional definitions require piloting.

Illustrative Application

Suppose a postgraduate researcher asks an AI tool to group studies on AI feedback in second language writing. The AI output may group the literature into "feedback quality", "learner engagement", and "writing improvement". This hypothetical example reports no observed findings. The categorisation may hide important distinctions. Some studies may examine learner perceptions rather than engagement. Others may focus on revision outcomes rather than writing development. A prompt-to-citation audit would require the researcher to record the original grouping prompt and output, identify each AI-generated category, check the category against the source articles, and document whether each article genuinely belongs in that category.

For example, treating improved revisions as evidence of lasting writing development would be overgeneralised; describing a perceptions questionnaire as behavioural engagement evidence would be misleading. Other potential risks include citation inflation, where citations are attached to claims they do not support; conceptual smoothing, where theoretical differences are flattened into neat categories; cross-language distortion; and citation fabrication, where references identify non-existent works. The final synthesis should distinguish perceptions, revision, and development, citing each source only for the outcome it actually examines.

Discussion and Implications

The framework provides a practical record of how researchers check and transform AI output into scholarly prose. This is especially relevant to applied linguistics because literature synthesis often shapes theoretical framing, research gaps, and methodological justification in areas such as L2 writing, feedback literacy, multilingual education, and technology-mediated language learning.

The framework may be used in a research notebook, thesis appendix, or postgraduate methodology training. A full audit can be time-consuming,

especially for postgraduates working with large literatures, multilingual sources or complex theoretical claims. Researchers can reuse verified source notes and prepare detailed audit records for candidate manuscript claims, categories and cross-study comparisons, while archiving complete synthesis sessions. This selectivity concerns documentation effort: every adopted AI-assisted claim and reference still requires source checking. Unfamiliar sources, causal claims and cross-language interpretations warrant closer scrutiny. If workload becomes excessive, reduce the scope of AI-assisted synthesis.

The framework remains a methodological proposal. It does not replace systematic review protocols, guarantee accuracy, or remove dependence on researchers' source access and interpretive expertise. Pilot studies should test category consistency, error detection, and total time, including verification, compared with conventional manual synthesis. Any efficiency gain remains an empirical question. Final responsibility for citation accuracy, interpretation, and scholarly accountability remains with the researcher.

References

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, Article 100745. https://doi.org/10.1016/j.asw.2023.100745

Boell, S. K., & Cecez-Kecmanovic, D. (2014). A hermeneutic approach for conducting literature reviews and literature searches. *Communications of the Association for Information Systems, 34*, 257–286. https://doi.org/10.17705/1CAIS.03412

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research, 26*, Article e53164. https://doi.org/10.2196/53164

Clark, J., Barton, B., Albarqouni, L., Byambasuren, O., Jowsey, T., Keogh, J., Liang, T., Moro, C., O'Neill, H., & Jones, M. (2025). Generative artificial intelligence use in evidence synthesis: A systematic review. *Research Synthesis Methods, 16*(4), 601–619. https://doi.org/10.1017/rsm.2025.16

Elsevier. (2026, June). *Generative AI policies for journals*. Retrieved October 1, 2026, from https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039