2026, vol. 12, pp. 23-32
RESEARCH ARTICLE https://doi.org/10.5281/zenodo.21032799

PROMPTING AS METHOD: DISCLOSURE THRESHOLDS FOR AI-ASSISTED LITERATURE SYNTHESIS

Dan Valeriu Voinea

University of Craiova

Abstract

Large language models are now used throughout literature review workflows, including summarization, extraction support, comparison across studies, and manuscript preparation. Existing disclosure norms have responded mainly by asking authors to identify whether generative AI was used, which tool was used, and for what general purpose. These requirements are necessary, but they are insufficient when prompts and iterative model interactions contribute to the interpretive work of literature synthesis. This article develops a threshold-based framework for prompt disclosure in qualitative and mixed-methods research. Drawing on construct validity, qualitative audit-trail traditions, qualitative evidence synthesis, review-reporting standards, and recent AI disclosure guidance, it argues that prompts should be treated as methodological traces when they shape cross-paper relationships, explanatory claims, causal hierarchies, typologies, metaphors, or adopted interpretive frames. In such cases, disclosure should extend beyond the prompt text itself to include the interaction path, supplied context, model and settings where available, decision points, human evaluation criteria, and the rationale by which model-generated interpretations were accepted, rejected, or revised. Lighter disclosure remains appropriate for grammar editing, formatting, translation used only for comprehension, unused brainstorming, and descriptive summary that does not enter synthesis. The proposed checklist links disclosure burden to interpretive consequence rather than to AI use as such. It treats prompt disclosure as an auditability practice that strengthens construct-validity warrants in AI-mediated knowledge production, without promising exact reproducibility.

Keywords: AI-assisted literature synthesis, prompt disclosure, construct validity, audit trail, qualitative methods, large language models

1. Introduction

Literature synthesis is no longer confined to database searching, screening, extraction, appraisal, and human-written narrative integration. Researchers increasingly use large language models (LLMs) to summarize papers, compare findings, draft thematic maps, identify possible research gaps, and formulate explanations across bodies of scholarship. These uses are especially attractive in qualitative and mixed-methods research, where the labor of organizing concepts, naming patterns, and developing explanatory accounts is intensive and often difficult to make visible. Recent work on LLMs in scientific reviews describes their use across review automation, screening, extraction, analysis, and manuscript preparation (Luo et al., 2024; Scherbakov et al., 2025). Studies of AI-assisted qualitative analysis similarly describe LLMs as tools that can contribute to coding, theme development, deductive textual analysis, and interpretive sensemaking, while still requiring human judgment and contextual knowledge (Lee et al., 2024; Morgan, 2023; Tai et al., 2024). More broadly, communication scholarship has framed AI as transforming media systems rather than merely accelerating text production (Stănescu, 2026), a distinction that is also relevant to research practice.

This development creates a reporting problem. Medical, publisher, and research-integrity guidance increasingly asks authors to disclose whether AI systems were used, to identify the tools involved, and to clarify the purposes for which they were used (Centers for Disease Control and Prevention [CDC], 2026; Elsevier, n.d.; European Commission, 2026; Flanagin et al., 2024; International Committee of Medical Journal Editors [ICMJE], 2026; Wiley, 2025; Zielinski et al., 2024). These policies are important. They help prevent undisclosed AI-generated text, preserve authorial responsibility, and clarify that AI systems cannot be credited as authors. Yet a generic statement that an author used AI to assist with literature synthesis does not show how the synthesis was produced. It does not indicate whether the model merely summarized individual articles, proposed relations among studies, ranked explanatory mechanisms, or supplied a conceptual frame that later appeared in the manuscript.

The distinction matters because literature synthesis is not a neutral condensation of prior work. In qualitative and mixed-methods traditions, synthesis often constructs relationships among studies: which findings belong

together, which concepts travel across contexts, which mechanisms explain observed variation, and which gaps are theoretically consequential. Reporting standards such as PRISMA 2020, PRISMA-S, ENTREQ, eMERGe, and SRQR already assume that consequential methodological choices should be made visible to readers (France et al., 2019; O'Brien et al., 2014; Page et al., 2021; Rethlefsen et al., 2021; Tong et al., 2012). LLM use raises a specific question within that broader transparency regime: when do prompts become part of the method?

This article addresses that question by asking how prompts and iterative LLM interactions shape interpretive constructs in AI-assisted literature synthesis, and what disclosure standard qualitative and mixed-methods researchers should apply. The central argument is that prompts should be disclosed as methodological traces when they contribute to explanatory or construct-shaping work. Prompting is not inherently a method, a dataset, an instrument, or an analysis. Its methodological status depends on what the interaction does in a particular workflow. When an interaction affects the interpretive construct reported in the synthesis, it becomes part of the warrant for that construct and should be disclosed accordingly.

The contribution is a threshold-based disclosure framework. Rather than requiring authors to preserve and publish every prompt entered into a model, the framework distinguishes between low-stakes AI assistance and interactions that shape interpretation. Full prompt and interaction-path disclosure is warranted when LLM use contributes to cross-paper relationships, explanatory claims, causal hierarchies, typologies, metaphors, or adopted interpretive framings. Minimal disclosure is usually sufficient for grammar editing, formatting, translation used only for comprehension, unused brainstorming, and summary without synthesis. The aim is to make AI-mediated analytic decisions sufficiently inspectable for readers, reviewers, and editors, without compiling an exhaustive archive of machine interaction.

2. Disclosure, Validity, and Auditability

This article is a conceptual methods paper. It does not offer a systematic review of AI use in evidence synthesis, nor does it empirically test the effects of prompt variation on review conclusions. Instead, it develops a reporting framework from four bodies of methodological discussion. AI disclosure guidance sets the current policy baseline, and review-reporting standards identify the procedural choices that must be visible in literature synthesis. Qualitative evidence synthesis explains why synthesis is interpretive rather than merely aggregative, while construct validity and the qualitative audit-trail tradition supply the basis for treating consequential prompts as part of an interpretive warrant.

Current AI disclosure guidance is largely organized around authorship, accountability, and tool identification. ICMJE asks authors to disclose AI-assisted technologies used in manuscript preparation and to preserve transparency in the editorial process (ICMJE, 2026). JAMA Network guidance similarly emphasizes transparent, appropriate, and accountable reporting of AI use (Flanagin et al., 2024), while WAME links generative AI disclosure to editorial accountability and authenticity (Zielinski et al., 2024). Publisher guidance generally follows the same pattern by distinguishing acceptable forms of AI assistance from prohibited authorship claims and by requiring disclosure in at least some circumstances (Elsevier, n.d.; Wiley, 2025). More recent institutional and governmental sources extend the discussion into the research process itself. The CDC's 2026 resource addresses generative AI use in proposing, preparing, performing, and reviewing scientific products (CDC, 2026), and the European Commission's living guidance connects disclosure to reliability, honesty, confidentiality, and accountability for research outputs (European Commission, 2026).

These policies establish an important baseline, but they are not designed as qualitative methodology standards. They generally answer a first-order question: did the author use AI, and for what broad purpose? They do not consistently answer a second-order methodological question: did the AI interaction shape the analytic object that the paper reports? That omission is understandable. Cross-disciplinary publisher rules must cover uses ranging from spell-checking to data analysis. They cannot easily specify how a prompt used to polish prose differs from a prompt used to identify the central mechanism in a qualitative evidence synthesis.

Review methodology provides a more appropriate analogy. PRISMA 2020 asks systematic reviewers to report enough about search, selection, appraisal, and synthesis for readers to understand what was done and to judge the credibility of the review (Page et al., 2021), and PRISMA-S extends that demand to the literature search itself (Rethlefsen et al., 2021). ENTREQ and SRQR carry similar transparency expectations into qualitative synthesis and qualitative research reporting, emphasizing methodological choices, trustworthiness practices, and the movement from evidence to interpretation (O'Brien et al., 2014; Tong et al., 2012). eMERGe is especially relevant, since it treats the analytic work of meta-ethnography, including translation and synthesis, as reportable rather than invisible interpretive labor (France et al., 2019).

Qualitative evidence synthesis strengthens this point because it treats synthesis as transformation. In meta-ethnography, the analyst translates concepts across studies and builds a line of argument (Noblit & Hare, 1988). Thematic synthesis separates descriptive themes from the more abstract analytic themes built upon them (Thomas & Harden, 2008). Realist review organizes evidence through relationships among context, mechanism, and outcome (Pawson et al., 2005), while critical interpretive synthesis develops synthetic constructs rather than simply aggregating findings (Dixon-Woods et al., 2006) and qualitative metasynthesis treats synthesis as an interpretive transformation of its sources (Sandelowski & Barroso, 2007). If an LLM participates in this work, whether translating across studies, developing analytic themes, constructing mechanisms, or naming a synthetic construct, it has entered a domain that existing qualitative traditions already regard as methodologically consequential.

Construct validity offers a useful vocabulary for this problem, provided that its use is bounded. Cronbach and Meehl framed construct validation as a problem of interpreting observed responses as evidence for underlying constructs (Cronbach & Meehl, 1955), while Messick later developed validity as a unified concern with the evidential and consequential basis for interpretation and use (Messick, 1995). Literature synthesis is not psychological measurement, and LLM prompts are not tests in any simple sense. The transferable point is narrower: validity concerns the warrant linking a procedure, a representation, and the construct that is finally reported. In AI-assisted synthesis, a researcher begins with a body of literature and ends with an interpretive construct, such as a typology, mechanism, explanatory frame, or line of argument. If an LLM interaction contributed materially to that construct, readers need enough information to evaluate the mediated path by which the interpretation emerged.

The qualitative audit-trail tradition supplies the practical counterpart to this validity concern. Lincoln and Guba's account of trustworthiness made dependability and confirmability central to qualitative inquiry, and audit practices became one way to document the movement from data to interpretation (Lincoln & Guba, 1985). Carcary's discussion of research audit trails is especially useful because it treats the audit trail as physical and intellectual documentation of key methodological and analytic decisions (Carcary, 2020). An audit trail does not require researchers to record every thought or abandoned possibility. It asks them to preserve the decisions that matter for understanding how the analysis developed.

This logic translates well to LLM-assisted synthesis. A prompt log is useful only if it clarifies how a model-mediated interaction affected the interpretation. A complete transcript of every exchange may be burdensome, noisy, ethically problematic, and methodologically uninformative. It may include false starts, irrelevant outputs, copyrighted material, confidential notes, or discarded ideas. Conversely, a short record of consequential interactions can make the analytic path inspectable without turning disclosure into indiscriminate surveillance. The relevant object is therefore not the prompt as a free-standing text, but the interaction path around a consequential decision: what the model was asked to do, what context it received, how it responded, how the researcher evaluated the response, and what interpretive decision followed.

LLM-specific reporting guidance is beginning to move in this direction. TRIPOD-LLM, for example, asks authors of LLM prediction-model studies to report details about prompts, model configuration, data, and evaluation so that readers can understand the study design (Gallifant et al., 2025). Its domain is not qualitative literature synthesis, but its reporting logic is relevant. Once an LLM becomes part of the research procedure, a tool name is not enough. Readers need to know what task the model performed, what information it received, and how its output was judged.

Prompt sensitivity research gives this issue further methodological weight. Technical work on prompting treats prompts as task-conditioning devices that shape model behavior (Liu et al., 2023). Sclar et al. (2024) show that LLM outputs can be highly sensitive to seemingly minor prompt-formatting choices in few-shot settings. Their findings do not directly establish how prompts affect qualitative synthesis, but they demonstrate that prompt design can materially alter model output. In literature synthesis, such differences can affect more than wording. They may elicit different explanatory categories, different causal priorities, or different views of which studies belong together. When that happens, prompting has ceased to be an interface detail and has become part of the method by which interpretation was produced.

3. When Prompting Becomes Methodologically Consequential

The methodological status of a prompt depends on its role in the research workflow. A prompt used to correct grammar or convert references into a journal style may require disclosure under publisher policy, but it does not normally shape the construct reported by the synthesis. A prompt that asks which mechanism best explains divergent findings across a set of studies is another matter. If the resulting mechanism is accepted,

revised, and incorporated into the argument, the prompt interaction has contributed to the interpretive warrant.

The threshold proposed here is explanatory or construct-shaping use. An LLM interaction crosses that threshold when it affects one of four elements of the final synthesis: relationships across papers, explanatory claims, causal or mechanism hierarchies, or adopted interpretive framings. Cross-paper relationships include claims that certain studies belong together, that one body of work extends another, or that several findings instantiate a shared pattern. Explanatory claims include statements about why a pattern occurs, why studies diverge, or how a mechanism links concepts. Causal hierarchies include rankings of causes, mechanisms, barriers, or facilitators as primary, secondary, upstream, or downstream. Adopted interpretive framings include model-suggested labels, metaphors, typologies, or organizing concepts that appear in the final manuscript.

This threshold marks the difference between description and interpretive construction. Description reports that several studies discuss a theme, use a method, or identify a pattern. Interpretive construction explains why that pattern occurs, identifies which mechanism gives the literature its structure, or names the concept through which the literature should be understood. An LLM-assisted summary stating that five studies discuss audit trails remains descriptive if the researcher uses it only as a reading aid and verifies it against the sources. The picture changes once the model's output makes a claim about the literature itself. A synthesis asserting that audit trails convert interpretive choices into confirmability warrants has entered explanatory territory, and one that names prompting as a form of analytic decision logging has entered construct-shaping territory. Such claims may be entirely defensible. The issue is that readers should know whether a model-mediated interaction helped produce them.

Four diagnostic questions can help authors and reviewers identify consequential cases. Did the model introduce a relationship among papers that was absent from the researcher's prior notes? Did it name a mechanism, typology, metaphor, or label that survives into the manuscript? Did it affect the relative importance assigned to competing explanations? Did it lead the author to accept, reject, narrow, or reframe a claim about what the literature means? A positive answer does not invalidate the work. It places the interaction within the interpretive warrant and therefore within the scope of methodological disclosure.

This approach also clarifies what prompt disclosure should not try to achieve. Exact reproducibility is often unrealistic in LLM workflows. Models are updated, platforms change, retrieval layers vary, system prompts remain opaque, and outputs may differ even when user-visible prompts are similar. These limitations do not make reporting pointless. Search engines change, databases update, and human analysts do not reproduce interpretive work identically either, yet reporting remains useful because it allows others to inspect the conditions under which claims were produced. For AI-assisted synthesis, the stronger standard is auditability rather than exact output reproduction. Readers should be able to examine the consequential decisions that shaped the interpretation and judge whether the construct-validity warrant is plausible.

Treating prompts as analytic decision logs also prevents two errors. The first is under-disclosure: hiding model-mediated interpretation behind a generic statement such as "AI was used for synthesis assistance." The second is over-disclosure: requiring authors to publish every trivial exchange with a model. Neither approach serves methodological transparency. The relevant unit is the decision point at which an interaction affected the synthesis. That decision point may include an initial prompt, supplied context, model output, human critique, revised prompt, and final accept/reject decision. The full interaction path, more than the prompt wording alone, is what allows readers to see how interpretation was shaped.

Consider a researcher who has extracted notes from 30 studies on AI-assisted qualitative analysis and asks an LLM to identify the main reason studies disagree about trustworthiness. The first output proposes methodological heterogeneity. The researcher asks for counterexamples and source-specific support, after which the model revises the explanation toward inconsistent construct operationalization. The researcher checks this against the extraction notes, finds partial support, rejects the stronger claim that this is the dominant explanation, and writes a narrower statement: differences in how studies operationalize trustworthiness partly explain divergent assessments of AI-assisted coding. In this example, the prompt interaction changed the explanatory claim that appears in the synthesis. Full disclosure is warranted because the model entered the analytic path by which the final interpretation was produced.

The same logic distinguishes LLM decision logs from other audit objects. The existing records each capture something different. Search logs show how sources were identified and filtered, analytic memos track the researcher's evolving interpretation, codebook revisions record category development, and reflexive journals document how standpoint and judgment influenced analysis. LLM decision logs add one more kind of trace, recording the model-mediated interactions through which candidate interpretations were generated,

challenged, modified, or rejected. They do not replace these other records, but they mark the specific points where LLMs participate in interpretation.

4. A Threshold-Based Disclosure Framework

The proposed framework has two levels: minimal disclosure and full prompt/interaction disclosure. Its purpose is to align disclosure burden with interpretive consequence. It is not meant to build a compliance archive or to imply that all AI use has the same methodological status.

Minimal disclosure is appropriate when AI use does not shape explanatory claims, interpretive constructs, causal hierarchy, or adopted framing. It should identify the AI role, tool or model, model version if available, date or period of use, and any settings that materially affected the task. This level is usually sufficient for grammar editing, formatting, reference-style assistance, translation used only to aid comprehension, unused brainstorming, and summarization that remains descriptive and is not used as the evidentiary basis for synthesis. Minimal disclosure remains compatible with publisher policies that require transparency about AI involvement while distinguishing writing support from research-method use (Elsevier, n.d.; ICMJE, 2026; Wiley, 2025).

Full prompt and interaction disclosure is warranted when an LLM interaction contributes to any of the four construct-shaping functions identified above: cross-paper relationships, explanatory claims, causal or mechanism hierarchies, or adopted interpretive framings. In these cases, authors should report enough for readers to understand the model's role in the analytic path. Disclosure should include the task framing, the context supplied to the model, the consequential prompt and revision path, the model and settings where available, the human evaluation regime, and the interpretive decision record. Where platform settings or model details are unavailable, authors should state that they were unavailable rather than supplying guessed values.

The framework can be summarized as follows:

LLM use	Appropriate disclosure	Rationale
Grammar, formatting, reference-style assistance, or copyediting	Minimal disclosure	The interaction affects expression rather than the interpretive construct, unless substantive meaning changes.
Translation used only for comprehension and checked by the researcher	Minimal disclosure, with source-verification procedures if relevant	The translation supports reading but does not itself generate synthesis.
Unused brainstorming	Minimal disclosure if required by venue policy	Discarded ideas do not enter the analytic warrant.
Descriptive summary of individual sources	Minimal disclosure, plus verification procedures when summaries inform screening or extraction	Summary becomes higher stakes when it substitutes for direct engagement with sources.
Cross-paper comparison, explanation generation, mechanism ranking, typology development, metaphor selection, or adopted conceptual framing	Full prompt and interaction disclosure	The model contributes to the interpretive construct reported in the manuscript.
Any case where a reader might evaluate the claim differently after seeing the prompt path	Full prompt and interaction disclosure	The interaction bears on the credibility of the reported interpretation.

For consequential interactions, the disclosure record should be selective but complete enough to support auditability. A useful record includes six elements. First, task framing: what the model was asked to do. Second, context selection: which sources, abstracts, excerpts, extraction notes, memos, or examples were supplied. Third, prompt and revision path: the initial prompt and consequential revisions that affected interpretation. Fourth, model and settings: model name, version, date, platform, temperature or comparable setting if available, retrieval mode if relevant, and known platform constraints. Fifth, human evaluation: how the

researcher judged, challenged, verified, rejected, or accepted the output. Sixth, interpretive decision record: which model-mediated decision points changed the final synthesis.

A methods statement need not reproduce every raw exchange. It should identify the consequential interaction and show how the author evaluated it. In the example above, a suitable disclosure might read:

LLM assistance was used after human extraction and memoing to generate candidate explanations for disagreement across included studies. The model was supplied with article identifiers, abstracts, extraction-note categories, and preliminary author memos. Two consequential prompt iterations are documented in Appendix A. The first suggested methodological heterogeneity as the main explanation; after the author requested counterexamples and source-specific support, the second suggested inconsistent construct operationalization. The final manuscript adopts a narrowed version of the second explanation after manual checking against extraction notes. Protected source text is not reproduced; Appendix A reports prompt templates, context categories, output summaries, and author evaluation notes.

A minimal disclosure statement would be shorter and would draw a boundary around non-interpretive use:

AI assistance was used for grammar editing and formatting of author-written prose. No AI output was used to generate, select, or revise the literature synthesis, explanatory claims, interpretive constructs, causal hierarchy, or conceptual framework.

Safe disclosure is sometimes necessary. Literature synthesis may involve copyrighted abstracts, unpublished notes, peer-review material, proprietary databases, or confidential data. Full disclosure should not require authors to release protected material. Instead, authors can report prompt templates rather than prompts containing copyrighted text, describe context categories rather than reproduce confidential excerpts, summarize model outputs rather than quote long passages, redact identifying details, or place sensitive appendices under controlled access when journals permit it. Safe disclosure must still preserve the analytic warrant. If redaction prevents readers from seeing how an interaction shaped the interpretation, the disclosure is too thin. If disclosure exposes protected material irrelevant to the decision, it is too broad.

The framework also gives reviewers and editors a basis for restraint. Reviewers should not demand complete prompt transcripts simply because a manuscript mentions AI use. They should first ask whether AI-mediated interaction plausibly affected an interpretive claim under review. If the disclosed role was grammar editing, formatting, translation for comprehension, or unused brainstorming, full prompt logs are usually unnecessary. If the manuscript contains a substantial explanation, causal hierarchy, typology, or conceptual frame, and the AI disclosure states that LLMs assisted with synthesis, reviewers can reasonably request the consequential prompt path and evaluation record. Editors can support this distinction through submission forms that separate writing assistance from data analysis, coding, synthesis, explanation generation, and conceptual framing.

The framework is best introduced as developmental rather than punitive. Many researchers do not yet have infrastructure for preserving AI interaction traces, and norms around prompt disclosure remain unsettled. A developmental approach would encourage prospective decision logging, ask reviewers to focus on validity rather than surveillance, and give authors a defensible basis for distinguishing consequential from trivial AI use. The goal is methodological accountability, not bureaucratic accumulation.

5. Boundary Conditions, Objections, and Future Work

Several recurring boundary cases illustrate why a threshold approach is preferable to a blanket rule. Summarization is the most common. Asking an LLM to summarize a single article, then checking that summary against the source and using it as a reading aid, usually falls under minimal disclosure. The consequential act remains the researcher's own engagement with the source. The situation changes when the model summarizes a large set of articles and those summaries become the evidentiary basis for cross-paper interpretation. At that point, the prompt template, context selection, summary-checking process, and source-verification procedure belong in the methodological record.

Translation presents a similar distinction. It is relatively low stakes when used only to support comprehension of a source that the researcher evaluates independently. It becomes methodologically consequential when translated passages are used to generate themes, compare concepts, or support explanatory claims across multilingual literatures. Concepts do not always map cleanly across languages, and an LLM translation may stabilize one interpretation while obscuring another. In such cases, the translation interaction bears directly on construct validity and should be disclosed in more detail.

Brainstorming should generally remain low burden. Researchers should be able to test ideas, discard weak formulations, and explore alternatives without archiving every unused suggestion. The exception arises when a brainstormed term becomes structurally important to the final argument. If an LLM suggests a label such as

analytic decision logging and that label becomes the central construct of the manuscript, the interaction is reportable. The reason is not that the model first produced the words; it is that the phrase became part of the paper's interpretive architecture.

Hallucination checking is related but distinct. Factual accuracy remains essential, and model outputs should be verified against source material. Yet hallucination is not the only methodological risk. A model may produce an accurate and well-supported output that nevertheless steers a synthesis toward one explanatory frame over another. Source verification and interpretive disclosure therefore serve different purposes. Verification addresses factual reliability. Disclosure of consequential prompt interactions addresses the credibility of the interpretive path.

The framework also faces several objections. The first is subjectivity. Authors may disagree about which interactions changed an explanation or construct, and some may be tempted to classify consequential interactions as minor. This risk is real. The framework cannot eliminate judgment, but it can make the judgment contestable by anchoring it to observable functions: cross-paper relationships, explanatory claims, causal hierarchy, and adopted framing. A reviewer can ask where a typology, mechanism, or conceptual label came from rather than issuing a vague demand for all prompts.

A second objection is conceptual overreach. Construct validity originates in measurement and assessment, not in every form of qualitative interpretation. The argument here does not treat prompts as measurement instruments or literature reviews as psychological tests. It uses construct-validity reasoning in a more limited way: when prompts help elicit, stabilize, or revise the construct reported by the synthesis, the warrant connecting procedure to interpretation becomes relevant (Cronbach & Meehl, 1955; Messick, 1995).

A third objection is performative transparency. A long prompt appendix can appear rigorous while failing to show why the author trusted the output. This is why the framework emphasizes decision points, human evaluation criteria, and author rationale. A transcript without an interpretive decision record is weak disclosure. A shorter trace that shows how an explanation was generated, challenged, checked against sources, narrowed, or rejected is stronger. Prompt disclosure is not a substitute for source verification, critical synthesis, or reflexivity. It is a record that helps readers evaluate those practices.

Confidentiality and intellectual property create practical constraints. Authors may not be able to reproduce all supplied context, particularly when it includes copyrighted abstracts, unpublished extraction notes, proprietary database content, or confidential material. The framework does not override these constraints. It requires authors to describe the method of using such material, the categories of context supplied, and the decisions affected, while redacting or summarizing protected content where necessary. This is consistent with broader research-reporting practice: restricted data may remain restricted, but the analytic procedure should still be described.

Model drift is a further limitation. A disclosed prompt may not reproduce the original output because the model, platform, retrieval layer, safety policy, or system prompt has changed. The limitation should be acknowledged rather than minimized. Nonetheless, reporting remains useful because it identifies the decision conditions under which a claim was produced. For AI-assisted literature synthesis, prompt disclosure should be understood as a condition-of-interpretation record, not as a guarantee of exact repetition.

The final limitation is empirical. The framework has not yet been tested. Future studies should examine how often prompt variation changes explanatory claims or adopted frames rather than merely altering wording. They should also test whether reviewers can reliably distinguish construct-shaping interactions from minor assistance, and whether the proposed disclosure levels are feasible in ordinary research workflows. Comparative studies could ask human-only and AI-assisted teams to synthesize the same literature and then compare the decision traces each produces. The aim should be to identify which decision points require documentation in each mode of work, rather than to rank AI-assisted and human-only synthesis against each other.

Such studies should avoid treating prompt sensitivity as a defect in isolation. Human analysts also vary in how they frame a literature, prioritize mechanisms, and name constructs. The relevant methodological question is whether the interpretive variation is visible enough for readers to evaluate. AI-assisted synthesis requires disclosure norms because LLMs can participate in interpretation while leaving few conventional traces. A threshold-based framework offers one way to make those traces visible without requiring exhaustive prompt archives.

6. Conclusion

Prompt disclosure in AI-assisted literature synthesis should be governed by interpretive consequence rather than by tool novelty. When LLMs are used for grammar editing, formatting, translation for comprehension, unused brainstorming, or descriptive summary that does not enter synthesis, minimal disclosure is usually sufficient. When prompts and iterative interactions shape cross-paper relationships, explanatory claims, causal hierarchies, typologies, metaphors, or adopted interpretive framings, those interactions become methodological traces.

The framework developed in this article links AI disclosure to construct validity, qualitative synthesis, and audit-trail reasoning. It treats consequential prompts as analytic decision logs when they contribute to explanation or construct formation. This position avoids two unhelpful extremes: hiding model-mediated interpretation behind generic AI-use statements, and demanding exhaustive prompt transcripts for every minor instance of AI assistance. Qualitative and mixed-methods researchers need a disclosure standard that is methodologically serious but practically usable. A threshold-based checklist provides that standard by tying disclosure burden to the role of the model in producing interpretation.

The broader issue is more than technical. As AI systems become routine participants in reading, comparing, and explaining literatures, credible scholarship will depend on whether researchers can show how interpretation was produced. Accountable selectivity offers a workable norm: disclose the interactions that changed interpretation, explain how they were evaluated, protect material that should not be public, and treat minor tool use through lighter reporting. Such a standard is modest, but it makes AI-mediated interpretation reviewable.

7. AI Disclosure

During the preparation of this revised draft, the author used Claude (Opus 4.8, Anthropic) and OpenAI ChatGPT to assist with research-question refinement, structural revision, language editing, and citation cross-checking. These tools were not used to generate data, conduct empirical analysis, or originate the paper's scholarly claims. The author reviewed and edited the content and accepts full responsibility for the accuracy, originality, and integrity of the final manuscript.

8. References

Carcary, M. (2020). The research audit trail: Methodological guidance for application in practice. Electronic Journal of Business Research Methods, 18(2), 166-177. https://doi.org/10.34190/JBRM.18.2.008

Centers for Disease Control and Prevention. (2026, May 28). Considerations for disclosing generative AI use in scientific work. https://www.cdc.gov/ai/resources/considerations-for-generative-ai-use-in-scientific-work.html

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302. https://doi.org/10.1037/h0040957

Dixon-Woods, M., Cavers, D., Agarwal, S., Annandale, E., Arthur, A., Harvey, J., Hsu, R., Katbamna, S., Olsen, R., Smith, L., Riley, R., & Sutton, A. J. (2006). Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. BMC Medical Research Methodology, 6, 35. https://doi.org/10.1186/1471-2288-6-35

Elsevier. (n.d.). Generative AI policies for journals. Retrieved June 28, 2026, from https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals

European Commission, Directorate-General for Research and Innovation. (2026). Living guidelines on the responsible use of generative AI in research (3rd version). https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf

Flanagin, A., Pirracchio, R., Khera, R., Berkwits, M., Hswen, Y., & Bibbins-Domingo, K. (2024). Reporting use of AI in research and scholarly publication: JAMA Network guidance. JAMA, 331(13), 1096-1098. https://doi.org/10.1001/jama.2024.3471

France, E. F., Cunningham, M., Ring, N., Uny, I., Duncan, E. A. S., Jepson, R. G., Maxwell, M., Roberts, R. J., Turley, R. L., Booth, A., Britten, N., Flemming, K., Gallagher, I., Garside, R., Hannes, K., Lewin, S., Noblit, G. W., Pope, C., Thomas, J., … Noyes, J. (2019). Improving reporting of meta-ethnography: The eMERGe reporting guidance. BMC Medical Research Methodology, 19, 25. https://doi.org/10.1186/s12874-018-0600-0

Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., Demner-Fushman, D., Dligach, D., Daneshjou, R., Fernandes, C., Hansen, L. H., Landman, A., Lehmann, L., McCoy, L. G., Miller, T.,

Moreno, A., Munch, N., Restrepo, D., Savova, G., … Bitterman, D. S. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. Nature Medicine, 31(1), 60-69. https://doi.org/10.1038/s41591-024-03425-5

International Committee of Medical Journal Editors. (2026). Use of artificial intelligence in publishing. https://www.icmje.org/recommendations/browse/artificial-intelligence/

Lee, V. V., van der Lubbe, S. C. C., Goh, L. H., & Valderas, J. M. (2024). Harnessing ChatGPT for thematic analysis: Are we ready? Journal of Medical Internet Research, 26, e54974. https://doi.org/10.2196/54974

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

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), Article 195. https://doi.org/10.1145/3560815

Luo, X., Chen, F., Zhu, D., Wang, L., Wang, Z., Liu, H., Lyu, M., Wang, Y., Wang, Q., & Chen, Y. (2024). Potential roles of large language models in the production of systematic reviews and meta-analyses. Journal of Medical Internet Research, 26, e56780. https://doi.org/10.2196/56780

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. International Journal of Qualitative Methods, 22, Article 16094069231211248. https://doi.org/10.1177/16094069231211248

Noblit, G. W., & Hare, R. D. (1988). Meta-ethnography: Synthesizing qualitative studies. Sage. https://doi.org/10.4135/9781412985000

O'Brien, B. C., Harris, I. B., Beckman, T. J., Reed, D. A., & Cook, D. A. (2014). Standards for reporting qualitative research: A synthesis of recommendations. Academic Medicine, 89(9), 1245-1251. https://doi.org/10.1097/ACM.0000000000000388

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review: A new method of systematic review designed for complex policy interventions. Journal of Health Services Research & Policy, 10(Suppl 1), 21-34. https://doi.org/10.1258/1355819054308530

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Systematic Reviews, 10, 39. https://doi.org/10.1186/s13643-020-01542-z

Sandelowski, M., & Barroso, J. (2007). Handbook for synthesizing qualitative research. Springer Publishing Company.

Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A., & Lenert, L. A. (2025). The emergence of large language models as tools in literature reviews: A large language model-assisted systematic review. Journal of the American Medical Informatics Association, 32(6), 1071-1086. https://doi.org/10.1093/jamia/ocaf063

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=RIu5lyNXjT

Stănescu, G. C. (2026). Artificial intelligence and the transformation of the media system. Encyclopedia, 6(2), 45. https://doi.org/10.3390/encyclopedia6020045

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. International Journal of Qualitative Methods, 23, Article 16094069241231168. https://doi.org/10.1177/16094069241231168

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8, 45. https://doi.org/10.1186/1471-2288-8-45

Tong, A., Flemming, K., McInnes, E., Oliver, S., & Craig, J. (2012). Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Medical Research Methodology, 12, 181. https://doi.org/10.1186/1471-2288-12-181

Wiley. (2025). AI guidelines for researchers. Retrieved June 28, 2026, from https://www.wiley.com/en-us/publish/article/ai-guidelines/

Zielinski, C., Winker, M. A., Aggarwal, R., Ferris, L. E., Heinemann, M., Lapeña, J. F., Pai, S. A., Ing, E., Citrome, L., Alam, M., Voight, M., Habibzadeh, F., & WAME Board. (2024). Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. Current Medical Research and Opinion, 40(1), 11-13. https://doi.org/10.1080/03007995.2023.2286102