

Auditing LLM-mediated interpretation through source-to-claim documentation

Chongwon Park

cpark2@d.umn.edu

University of Minnesota

Bo Kyoung Kim

Georgia Institute of Technology

Article

Keywords: generative artificial intelligence, qualitative research, audit trail, reflexivity, reanalysis, meaning representation

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10756391/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Auditing LLM-mediated interpretation through source-to-claim documentation

Abstract

Generative artificial intelligence is used in qualitative coding and analytic writing, yet hosted models complicate auditability because repeated calls may vary and fluent prose can obscure how evidence becomes an analytical claim. Source-to-claim documentation is introduced as a framework that links source evidence with model interactions and review, then preserves the chain for reanalysis. Its contribution lies in retaining the evidential path through which generated material acquires analytical status, rather than disclosing that AI was used. A diagnostic demonstration with twelve constructed sources and two workflow conditions illustrates the framework's function. Repeated paraphrases reproduced the same semantic shift, while quotation-compliant memos introduced disputed interpretations. Average graph overlap also concealed item-level variation. Structured comparison can identify where a source relation changes, but human judgement remains responsible for deciding whether the change affects an analytical claim. The framework offers a proportional basis for auditing LLM-mediated interpretation in humanities and social research.

Keywords: generative artificial intelligence; qualitative research; audit trail; reflexivity; reanalysis; meaning representation

1 Introduction

Large language models now assist qualitative researchers across the humanities and social sciences with coding and analytic writing. Early studies point to gains in speed, but they disagree about the depth and reliability of machine interpretation. Morgan (2023) found that ChatGPT recovered concrete themes more readily than subtle ones. By contrast, Dunivin (2025) developed a reflexive approach in which model assistance remains embedded within human coding practice. Other work emphasises epistemological limits or cautious adoption rather than efficiency alone (Anis and French, 2023; Messner et al., 2025). Recent articles in the *International Journal of Social Research Methodology* extend the discussion through technological reflexivity and new forms of human-AI collaboration (Paulus and Lester, 2024; Phan and Le, 2026; Lai and Wu, 2026).

Across the literature, one difficulty recurs. An LLM can turn source material into fluent analytical prose within seconds, but the finished paragraph hides the transformations that occurred during generation. A reader may encounter an interpretation without knowing whether a commitment came from the source or arose during machine generation and human revision. Disclosure cannot resolve the problem by itself. Naming the AI system identifies a tool, but it does not reconstruct how a generated artefact influenced a claim.

A second difficulty arises from the use of hosted systems. Researchers do not control a fixed computational object, as a provider may retain a public model name while changing the underlying service. Repeated calls can also yield different wording under apparently stable conditions. Exact regeneration is uncertain, but output identity would remain a poor standard even if it were available. Different responses may preserve the same evidential relation, whereas repeated responses can reproduce the same distortion.

Qualitative research already offers a methodological resource for addressing such mediation. Audit trails preserve the decisions through which data become interpretation, and reflexive accounts explain how the researcher and setting shape analysis (Lincoln and

Guba, 1985; Rodgers and Cowles, 1993; Berger, 2015). Digital workflows widen the obligation because software participates in the transformation of evidence (Paulus and Lester, 2024). When a generative system contributes new language, its output should remain available beside the human decision that follows.

The source-to-claim documentation framework proposed here extends audit trails for interpretive work in the humanities and social sciences. It links the research frame and source to a dated generation record, followed by human review. The phrase “source-to-claim” denotes the evidential path from original material to the analytical claim eventually reported. When a claim depends on a semantic relation that fluent prose may alter, the method can add a structured checkpoint, understood here as a compact representation placed alongside the generated text. In the diagnostic demonstration, candidate meaning graphs served that function. A meaning graph represents selected aspects of a sentence, such as event structure or evidential status, through labelled relations among concepts. By making the audited relation visible and comparable across runs, the graph helps reviewers identify where a model preserved or changed the source commitment. The broader method does not require specific graph notation because another structured record can serve the same purpose when it keeps the relevant relation inspectable.

The central question concerns how qualitative audit trails should change when a hosted generative model mediates the movement from evidence to analytical claim. Source-to-claim documentation addresses the problem by preserving model output and human review as linked research objects. Within the broader framework, the graph audit asks whether repeated outputs preserve a relation identified in the source rather than merely reproduce similar wording. A diagnostic demonstration illustrates how the general framework and graph-based procedure operate without treating the results as estimates of model performance.

2 Audit trails and technological reflexivity

2.1 *Trustworthiness without mechanical replication*

Qualitative audit trails were developed to substantiate trustworthiness through records of context and analytical decision-making. Rodgers and Cowles (1993) describe the trail as a body of documentation rather than a single log, since methodological notes and reflexive records answer different questions about an inquiry. Lincoln and Guba (1985) connect such records to dependability and confirmability. Later guidance continues to favour systematic disclosure without reducing quality to procedural uniformity (Nowell et al., 2017; Tracy, 2010; Levitt et al., 2018).

Debates over transparency reveal an unresolved tension in qualitative methodology. Aguinis and Solarino (2019) call for reporting practices that permit closer inspection and, where appropriate, replication, whereas Pratt et al. (2020) caution that transparency loses its value when it is equated with reproducing the same interpretation. The source-to-claim documentation proposed here addresses the tension by separating evidential inspectability from interpretive agreement. It preserves the path through which source material became an analytical claim, allowing another scholar to examine the transformation without being expected to endorse the resulting reading.

The advantage of the source-to-claim documentation framework becomes clearer when repeatability is separated from reanalysis. Repeatability asks whether archived code can regenerate a reported summary from fixed files, which remains central to computational reproducibility (Sandve et al., 2013; National Academies of Sciences, Engineering, and Medicine, 2019). Reanalysis has a different aim because it asks whether preserved research objects can support a later reading under another analytical frame (Schöch, 2023). Qualitative data reuse requires contextual records, since detached transcripts cannot reveal the conditions under which an earlier interpretation was formed (Bishop, 2007; Irwin, 2013). Source-to-claim documentation supports both activities by preserving fixed computational artefacts together

with the context and review decisions needed for reinterpretation. Such continuity is especially important in LLM-mediated research because the hosted model may change even though its dated output continues to shape a published claim.

2.2 Generative systems as methodological participants

Computer-assisted analysis has long influenced qualitative workflows, including the development and revision of theory (Dalkin et al., 2021). Generative systems differ from conventional software because they can select evidence and propose interpretation within one fluent response. A coding platform usually makes its operations visible through an observable workflow. An LLM can compress several analytical moves into a paragraph that reads as if no transformation occurred.

Methodological reflexivity requires researchers to examine how their design choices shape the knowledge they produce.¹ Paulus and Lester (2024) extend that practice to digital research through the concept of technological reflexivity, which directs attention to how digital tools and environments alter the research process and its outcomes. Generative systems widen the issue because the tool produces language that may enter the analysis directly. Phan and Le (2026) describe AI as a research partner while preserving the researcher’s responsibility for selection and interpretation. Lai and Wu (2026) similarly place prompt refinement and generative iteration within an exploratory methodological framework.

Nonetheless, the partnership metaphor can still obscure asymmetry. Fluent output can invite readers to attribute meaning or communicative intent to a system that generates linguistic form without such grounding (Bender et al., 2021). Researchers can explain why a passage was selected and why a claim was accepted. Model documentation and data statements address related problems at the system level (Bender and Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019), but an individual project needs a narrower record because scholars rarely possess complete information about a proprietary service. Their practical obligation is to preserve what they could observe and how the output affected analysis.

Prompt structure can also manufacture apparent validity. Zheng et al. (2026) distinguish prompt-valid outputs from theory-valid data, finding that plausible structure may reflect instructions rather than autonomous inference. Interpretive research faces the same risk when fluency hides how strongly the prompt shaped a claim. Later readers need the instruction beside the output if they are to evaluate the transformation.

Recent approaches address either reporting or process traceability. Jones’s TROUT-AI and Fehring et al.’s COREQ+LLM initiative strengthen the reporting of prompts and human oversight (Jones, 2025; Fehring et al., 2025). Lu et al.’s QualAnalyzer preserves the complete interaction for each source segment, whereas Forster et al. trace concepts back through a task-specific qualitative content analysis pipeline (Lu et al., 2026; Forster et al., 2026). Source-to-claim documentation draws the two strands together but serves a different function. Its four records preserve how an individual generated artefact moves from source evidence to analytical status, and an optional checkpoint makes a selected transformation available for review. The chain remains open to reanalysis across interpretive methods. Table 1 compares existing approaches with the source-to-claim documentation framework and shows how the framework extends reporting and traceability through a linked evidential path.

Table 1. Methodological functions of related approaches

Approach	Primary function	Relation to source-to-claim documentation
TROUT-AI and COREQ+LLM (Jones, 2025; Fehring et al., 2025)	Strengthen disclosure of AI use and human oversight.	Centre on reporting rather than a claim-level archive of linked artefacts.
QualAnalyzer (Lu et al., 2026)	Preserves the complete model interaction for each source segment.	Provides item-level traceability without a separate checkpoint linked to interpretive review.
Forster et al.’s pipeline (Forster et al., 2026)	Traces concepts back to interview passages within qualitative content analysis.	Offers a task-specific workflow rather than a representation-neutral framework for interpretive claims.
Source-to-claim documentation	Links each source to generation and review, then preserves the chain in an archive.	Shows how a generated artefact acquired analytical status and supports later reanalysis.

The distinctive contribution lies in the linkage rather than any single component. Reporting schemes can show that AI participated, while traceable pipelines can recover model-derived outputs. Source-to-claim documentation also preserves how an artefact acquired analytical status, allowing later inspection without requiring the original interpretation to be reproduced.

To preserve a clear division of responsibility, source-to-claim documentation treats the model as a dated computational instrument rather than an autonomous interpreter. Within that arrangement, the research team defines the task and evaluates the result, while the system contributes candidate artefacts for review. Reflexivity thus extends beyond the researcher’s position to include the technological mediation through which evidence is transformed. A defensible record should accordingly show how the model entered the workflow and how scholarly judgement governed the transition from generated output to analytical claim.

3 Source-to-claim documentation

The source-to-claim documentation framework consists of four linked records that preserve the path from source evidence to an analytical claim. It first documents the research frame and source together with the model interaction and raw output. It then preserves any structured audit and human review before linking the analysis to an archive that supports later inspection or reanalysis. Table 2 specifies the minimum evidence required at each stage. The framework remains compact because full annotation of every passage would be impractical for most interpretive studies.

Table 2. The source-to-claim record

Record	Function	Minimum evidence
Research frame and source	Defines the analytical claim and anchors it in recoverable material.	Task boundary and source identifier.
Model interaction	Records the computational instrument as it was available during the study.	Exact prompt and dated interaction record, including the raw output.
Structured audit and review	Makes selected transformations inspectable and records the decision that followed.	Candidate representation or comparison note, followed by a review rationale.
Analysis and archive	Supports verification over fixed files and later reinterpretation.	Analysis script and linked repository record with stable identifiers.

The following subsections describe each of the four linked records and explain the methodological function that each performs.

3.1 *Research frame and source*

The first record establishes the basis on which a generated artefact can be evaluated. It joins the research frame with the source because an output cannot be interpreted apart from the task assigned to the model or the evidence supplied to it.

The research frame specifies the analytical purpose of the model interaction. It should indicate whether the model was asked to summarise or interpret, while also defining the kinds of inference that the task permits. Such boundaries allow a reviewer to distinguish failure to follow the instruction from a legitimate disagreement about the source.

The source component identifies the passage and preserves its provenance. It should include the complete material shown to the model, since an output cannot be evaluated against a shorter excerpt than the one used during generation. The same principle applies when the prompt contains a codebook or prior examples. In each case, the evidence record must reflect the material that shaped the response.

The two components perform complementary functions. A quotation may remain accurate even when the surrounding memo exceeds the assigned task, so preserving the source alone cannot establish whether the output was methodologically acceptable. Conversely, a clear task description cannot support review unless the relevant evidence remains recoverable. When redistribution is restricted, a stable citation and documented access conditions can preserve the source-to-claim link without requiring public release.

3.2 Model interaction and raw output

The second record documents the computational event that produced the artefact. It links the complete prompt and observable model conditions to the unedited response, allowing a later reader to reconstruct what the system was asked to do and what it returned.

The prompt defines the evidence available to the model and the form of the requested response. An output preserved without its prompt loses the conditions needed for interpretation. Preserving the prompt alone creates a different gap, since later readers would have to call a hosted service that may no longer respond in the same way. The record should thus retain the exact instruction together with any contextual material supplied during generation.

A model manifest² situates the interaction within a specific computational context. Because researchers cannot inspect the hidden infrastructure of a hosted service, the manifest

cannot identify a permanent model snapshot. It instead preserves the conditions that were observable when the output was produced, following established documentation practices for data and models (Bender and Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019).

The unedited response completes the second record. It should be preserved before correction or selective quotation, since silent revision prevents readers from determining whether a commitment originated in the model or during later editing. Any reviewed version should remain linked to the original output, but the explanation for accepting or changing it belongs to the subsequent review record. When public release would expose protected source or participant information, stable identifiers and controlled access can preserve the connection without requiring unrestricted disclosure.

3.3 Structured audit records and human review

The third record documents how a generated artefact was examined before it acquired analytical status. It combines an optional structured audit with a record of human review. The structured component identifies a source-level transformation, while the review component explains whether that transformation affects the claim.

Direct comparison with the source may be sufficient when the generated text remains close to the evidence and the relevant relation is easy to identify. A structured audit becomes useful when fluent prose compresses or obscures a relation on which the analysis depends. Rather than representing the entire interpretation, the audit isolates a limited source commitment, such as the assignment of agency or the evidential status of a proposition. The representation may take the form of a compact table or a graph, depending on the risk being examined. Its purpose is diagnostic because it identifies where closer inspection is needed without deciding whether the output is acceptable.

Human review supplies the interpretive judgement that the structured record cannot provide. The reviewer returns to the source, considers the identified transformation, and records whether the generated artefact can support the analytical claim. When revision is

required, the rationale should explain what changed and why. Disagreement should remain visible whenever it affects the use of the artefact, since an unexplained final decision would conceal the interpretive boundary that the audit was intended to expose.

The two components should remain linked. A structured representation without a review rationale cannot explain the analytical significance of a difference, whereas a review note without supporting evidence is difficult for another scholar to assess. The record need not preserve a comprehensive account of the reviewer’s reasoning. It should preserve enough evidence to show how the generated artefact acquired its final role in the analysis.

3.4 Archived analysis and reanalysis

The fourth record preserves the analytical materials needed to verify a published result and revisit it under a different interpretive frame. It links fixed outputs and analysis files to the earlier source, generation, and review records, thereby maintaining continuity across the entire source-to-claim path.

Verification should operate on archived outputs rather than on fresh model calls because a new response would create a different research object. Scripts can regenerate reported counts or comparison summaries from the preserved files, while stable identifiers connect each result to its source and review decision. A brief README should explain how the files relate so that another scholar can reconstruct the workflow without relying on undocumented knowledge.

Archival organisation also supports reanalysis. Saving every file is insufficient when the relationships among them remain unclear. Version notes should show how a generated artefact entered the analysis, whether it was revised, and why it was later retained or excluded. Another researcher can then inspect the original claim before applying a different coding scheme or interpretive framework to the same evidence.

The record supports both technical and evidential verification. Technical verification examines whether archived code reproduces the reported analysis from fixed files. Evidential

verification asks whether the analytical claim remains supported after the source, prompt, generated artefact, and review decision are considered together. A workflow may succeed computationally while producing a weak interpretation. Conversely, a persuasive reading may depend on an undocumented transformation. Linked records allow another scholar to locate the source of either problem.

Archiving must also account for consent and access restrictions. Generated material may reproduce protected sources, while review records may contain participant information that cannot be released publicly. Controlled access or redaction can preserve accountability when unrestricted disclosure would violate ethical or legal conditions. We suggest that the archive should document such limits explicitly, since NIST guidance treats generative AI risk management as an ongoing process that includes periodic review (Autio et al., 2024).

4 Implementing the four-record framework in a diagnostic experiment

Section 3 defined source-to-claim documentation as a framework of four linked records, and the diagnostic experiment put that framework into practice by preserving a stable source identifier from model generation through the review and archiving stages. Generated artefacts then added a dated model label and task-run code, whereas reader data added an anonymised rater code. The resulting links allow an item-level finding to be traced through every transformation to the source sentence that initiated it.

The implementation limited the scope of each audit component rather than applying every form of review to every artefact. Twelve constructed sources entered three tasks, and two hosted OpenAI models completed two runs of each task, producing 144 raw artefacts. Candidate graph pairs supplied the structured audit, while reader judgements evaluated prose from the primary model. The study functions as a worked methodological demonstration. It examines whether the four records expose different weaknesses, but it does not compare the framework with another audit procedure or measure gains in analytical decision-making.

4.1 *Record 1 in practice: research frame and source*

The first record established the purpose and scope of the experiment before any model call. Before generation began, the protocol specified that the constructed sources would be used to test whether the four linked records revealed different weaknesses in the workflow, rather than to estimate model accuracy or compare overall performance. More specifically, the protocol asked whether linked records could expose a source-level shift and distinguish it from later interpretive expansion. The examples were constructed for methodological control rather than realism. Each sentence isolated a relation that could be specified before generation, which made later movement easier to locate without inaccessible context. Katz et al. (2026) likewise used synthetic qualitative data to evaluate an LLM-assisted method under known conditions. Exact archiving also avoided licensing restrictions. The controlled setting carries a cost because the sources do not reproduce the density of an interview collection or documentary archive. Their role is to stress-test the links among records, not to represent the distribution of errors in humanities or social research data.

Table 3 presents the twelve constructed sources and explains how each pair isolates a semantic relation or interpretive boundary that may shift during generation. Twelve single-sentence sources were assigned identifiers S1 through S12 in the order shown in Table 3. Each source record paired the exact wording with the relation selected for audit. The attribution items focused on whether reported content remained attributed, while the remaining items tested another source relation or the boundary between literal content and pragmatic interpretation. The source manifest preserved the identifier and phenomenon label. A machine-readable stimuli file supplied the same sentence to every model condition, so later comparisons did not rely on a reconstructed source.

Table 3. Constructed diagnostic sources and their design rationale

Phenomenon	Source examples	Reason for inclusion
Attribution and evidential status	“A blogger alleged that the senator met with lobbyists, but the senator denied that any meeting occurred.” “The witness said the guard opened the gate after a phone call, and the judge noted that this testimony was hearsay.”	Tests whether reported content becomes an unqualified assertion or loses an evidential qualifier.
Counterfactual modality	“If the evacuation order had been issued earlier, the injuries would have been prevented.” “Had the team practiced more, they might have won the match, but they lost three to zero.”	Tests whether hypothetical events remain distinct from actual outcomes.
Partial negation and scope	“Not every reviewer supported the proposal; two reviewers abstained.” “The curator did not entirely dismiss the forgery claim, but she called the evidence weak.”	Tests whether limited denial is expanded into rejection of the whole proposition.
Reference resolution	“Carla criticized Dave during the meeting, and he apologized to her the next day.” “Rina lent her notes to Omar because he had missed the lecture.”	Tests whether participant links and event roles survive paraphrase.
Sarcasm as pragmatic stance	“After the system crashed again, Priya said, ‘Fantastic, just what we needed,’ and rolled her eyes.” “When the bus arrived thirty minutes late, Jamal muttered, ‘Right on time.’”	Tests whether literal wording is mistaken for the speaker’s intended evaluation.
Institutional metonymy	“London decided to cut the subsidy to the provinces.” “Washington announced new tariffs on steel imports.”	Tests whether place names are expanded into institutional actors without marking the inference.

Every prompt and output retained the source identifier, as did the comparison and reader records. Later artefacts could thus be evaluated against the task boundary and source relation established in Record 1.

4.2 Record 2 in practice: model interaction and raw output

The second record documented each computational event. For every call, the interaction record combined the source identifier with a dated model label and a task-run code. It also preserved the exact prompt and unedited response. Each JSONL entry also stored the observable metadata and completion status, allowing every artefact to be traced to the call that produced it.

Source S8 was “Rina lent her notes to Omar because he had missed the lecture.” In the primary GPT-4.1 condition, the source was paired with the following instructions. The structured instruction requested a PENMAN³ meaning graph informed by Abstract Meaning Representation (AMR)⁴ and Uniform Meaning Representation (UMR).⁵ PENMAN supplied the textual format used in the experiment.

Paraphrase prompt. Paraphrase the source sentence in one sentence. Do not add facts. Do not omit any named agents or targets. Preserve attribution (said, claimed, alleged, denied, noted), uncertainty, modality, and negation exactly. Return only the paraphrase. Do not use headings. Do not wrap your answer in code fences.

Structured-audit prompt. Produce a PENMAN meaning graph for the source sentence (AMR/UMR style). Represent attribution (allege, say, deny, note), modality/counterfactuality, negation scope, and reference explicitly when present. Keep agents explicit. Return only the PENMAN graph. Do not use headings. Do not wrap your answer in code fences.

Interpretive-memo prompt. Write a 120-word interpretive memo about agency and stance. Every claim must cite an exact phrase from the source sentence in straight double quotes like “this.” If a claim is not directly licensed by the text, label it as INFERENCE. Return only the memo. Do not use headings.

Both paraphrase runs produced the same response:

“Rina gave her notes to Omar since he had not attended the lecture.”

The repeated wording concealed a source-level change because *lend* preserves a temporary transfer and continued ownership, whereas *give* can imply a permanent transfer. The structured output supplied a separate record of the source relation. Its relevant content can be rendered without specialized notation:

event = lend
actor = Rina
resource = notes
owner = Rina
recipient = Omar
reason = Omar missed the lecture

The simplified rendering is not presented as a reference AMR or UMR annotation. It restates the relation encoded in the candidate graph, and another project could use a table or annotation record for the same purpose. The memo output moved farther from the source by describing Rina’s “motivation to help Omar” and her stance as “supportive or helpful.” The latter description was labelled as inference.

The primary artefact set was generated on 6 February 2026 with the system labelled ChatGPT GPT-4.1 in the study logs. A later workflow sensitivity condition used ChatGPT GPT-5.5 on 4 May 2026.⁶ Each condition produced 72 artefacts. The later condition was included to show how dated records remain interpretable as a hosted workflow changes, not to compare model performance. Temperature was set to 0.0, while top-p remained at 1.0.⁷

The primary collection began with ten sources. When the two metonymy items were added, the paraphrase and graph instructions were expanded to address reference and metonymy. The later GPT-5.5 condition used the resulting twelve-source protocol throughout. Cross-condition differences consequently reflect changes in both the hosted model and part of the prompting procedure rather than an isolated model effect. Every response was preserved without correction, including one graph-shaped artefact that could not be parsed.

4.3 Record 3 in practice: structured audit and human review

Once Record 2 had archived the unedited outputs, Record 3 compared and reviewed them. The source identifier and model label established each comparison pair, while the run number kept the two responses distinct. The comparison asked which source relation remained visible across repeated generation and where divergence warranted renewed inspection.

A script first checked whether each graph-shaped output could be read as a structured representation. An unreadable response remained in the archive as a generation failure rather than disappearing from the analysis.⁸ For readable pairs, normalisation reduced the influence of variable names and formatting. Smatch then supplied an F1 score for structural overlap between the two runs (Cai and Knight, 2013).⁹

Structural overlap served as a routing device rather than a judgement of accuracy. Two similar representations may repeat the same source error, whereas two defensible representations may use different structures. A low score identifies a pair that deserves inspection, while a high score shows only that the candidate records resemble one another. The source remains the standard against which either representation must be reviewed.

For S8, both graph outputs retained lending as the central event. They linked Rina to the notes and Omar to the recipient role. One run added temporal detail, but the relation selected for audit remained recoverable. Linking the graphs to the paraphrases identified where both prose runs departed from the source, although structural comparison alone could not determine whether the substitution affected an analytical claim.

Human review addressed that question. The human-participant component of the study was conducted entirely in the United States. Eight undergraduate students evaluated the GPT-4.1 prose for all twelve sources, yielding 96 source-level judgements.¹⁰ For each source, readers first compared the original sentence with two paraphrases and then evaluated two interpretive memos.

Paraphrase questions concerned fidelity and pairwise difference. Memo questions separated quotation compliance from interpretive support before asking whether the two memos differed in their readings. Two missing responses remained missing rather than being imputed. Each judgement retained the source identifier and an anonymised reader code. Readers did not evaluate graph notation because formal annotation expertise was outside the purpose of the survey.

Because both graph and prose were model-generated, the structured record did not

provide independent validation. Its role was to externalise a selected relation so that readers could return to the source. Human judgements then addressed whether the prose could support an analytical claim, and the two forms of evidence remained separate.

4.4 Record 4 in practice: archived analysis and reanalysis

Record 4 preserved the linked artefacts before any summary was calculated. The merged generation file stored one JSON object for every model response.¹¹ Each object joined the source identifier and generation condition to the stored prompt and unedited output.

The analysis scripts used archived identifiers rather than fresh model calls. Across the two workflow conditions, the study crossed twelve sources with three tasks and two runs, producing 144 generation records. The same identifiers formed 24 within-condition graph pairs. De-identified reader responses were joined to the primary prose artefacts through the source identifier and an anonymous reader code, allowing the reported summaries to be regenerated from fixed files.

The archive also preserved the development of the procedure. Exact instructions show which prompt version produced each artefact, allowing another analyst to reconstruct the workflow without treating the GPT-5.5 condition as a model-performance comparison.

The memo records support another form of reanalysis. For the allegation source, both GPT-4.1 memos quoted the allegation and denial but moved beyond the source when describing exposure or self-protection. A later study could recode the same outputs under a stricter distinction between source-licensed analysis and speculation without repeating the generation.

Record 4 consequently preserves more than a collection of files. It maintains the links between source evidence and the decisions based on generated artefacts. Technical verification can reproduce reported counts, while evidential reanalysis can revisit how an output acquired analytical significance.

4.5 A complete source-to-claim trace

Table 4 follows source S8 from the relation selected for audit to the review decision and archived analysis.

Table 4. The complete source-to-claim trace for source S8

Record	Archived evidence	Methodological function
Research frame and source	Source identifier S8_reference_he_because ; source text: “Rina lent her notes to Omar because he had missed the lecture.” The audit focus concerned the transfer event and participant links.	Establishes that the output should preserve lending, Rina’s relation to the notes, and Omar’s role as recipient.
Model interaction and raw output	The complete paraphrase prompt was stored with two dated GPT-4.1 responses. Both runs produced: “Rina gave her notes to Omar since he had not attended the lecture.” The graph and memo outputs retained their task and run identifiers.	Records what the model produced and shows that repeated generation introduced the same lexical substitution.
Structured audit and human review	Both candidate structures retained lend-01 , including Rina’s relation to the notes and Omar’s recipient role. Readers identified distortion in 62.5% of the paraphrase ratings, and 50% classified the change as substantial.	Identifies the difference between <i>lend</i> and <i>give</i> , then records whether readers regarded the change as analytically important.
Archived analysis and reanalysis	All linked artefacts and de-identified ratings retained the source identifier. Scripts regenerated the item-level summary from fixed files.	Allows another analyst to inspect or recode the finding without requesting a new model response.

The trace shows why agreement across repeated outputs cannot serve as a proxy for fidelity. Both paraphrases used the same wording, but their agreement concealed a departure from the source. The paired structures retained lending as the transfer event, while human review addressed whether replacing *lent* with *gave* affected the potential use of the paraphrase.

The memo branch exposed a different movement from source to claim. One output characterized Rina as “helpful or considerate,” while the other described her “motivation to help Omar” before labeling the broader characterization as inference. Because the archive preserves both branches, another analyst can revisit the transfer distinction or apply a different coding rule without altering the dated artefacts.

4.6 What the linked records revealed

The linked records exposed different weaknesses across the workflow, while the sarcasm items showed where structured audit reached its limit. Table 5 summarizes the principal illustrations.

Table 5. Representative findings from the linked records

Source or comparison	Observed pattern	Methodological implication
“Rina lent her notes to Omar”	Both primary-model paraphrase runs replaced <i>lent</i> with <i>gave</i> .	Agreement across runs concealed a shared source-level shift.
Allegation and tariffs sources	Memos quoted the source but introduced different claims about motive or policy stance.	Quotation compliance did not establish interpretive support.
Run-to-run graph pairs by workflow condition	Mean structural overlap was 0.84 in both conditions, although item-level overlap shifted.	Aggregate summaries obscured which items required renewed inspection.
Sarcasm sources	The graph records preserved the utterance and its incongruous context, while some also encoded a sarcasm label.	A structured label could expose an interpretation but could not validate pragmatic force without human review.

Paraphrases appeared relatively stable overall. Among 191 valid source-faithfulness ratings, 142 judged the paraphrase equivalent to the source, while 49 identified some distortion. Readers chose a tie in 90 of 96 preference comparisons, and only eight responses identified a meaning difference between the two runs. The lending trace shows why the aggregate pattern remains incomplete, since agreement concealed the same departure from the source.

Interpretive memos exposed another weakness. Readers found the required quotation in 191 of 192 judgements, yet they marked added interpretation in 125. Among 191 support ratings, 55 placed the memo in the mixed or unsupported range. The allegation example moved into different claims about exposure or self-protection, while the Washington tariffs item produced disagreement about the stance implied by the policy.

The later workflow sensitivity condition illustrated the value of item-level records. Ten

of twelve graph pairs in the primary condition differed after normalisation. The later condition contained one unparseable pair, leaving eleven pairs for within-condition comparison. Mean run-to-run F1 was 0.84 in both archived workflows, even though item-level overlap differed. The identical averages are descriptive rather than evidence of model equivalence because the second condition was not designed to assess model performance and the prompt history also changed.

The sarcasm sources established a boundary for structured audit. A graph could preserve the utterance and the context in which a late bus was described as “Right on time.” Some generated structures also inserted a sarcasm label, but the label already embodied an interpretation. Human review remained primary once the analytical question depended on pragmatic force.

In sum, the records form a sequence in which Record 1 defines the evidential boundary and Record 2 preserves generation. Record 3 documents evaluation, while Record 4 keeps the complete path available for verification or reinterpretation.

5 Methodological advantages and limits for interpretive research

5.1 Beyond disclosure and conventional audit trails

The comparison in Table 1 shows why disclosure alone is insufficient when generated material enters an analysis. Naming the AI system or describing its role cannot reveal which source produced a particular output, nor can it show how the output was evaluated.

Source-to-claim documentation connects the functions usually divided between qualitative audit trails and computational logs. A reviewer can move from an analytical claim to the reviewed artefact, then recover the raw output and source without being asked to reproduce the same interpretation. JARS-Qual likewise requires enough context and procedural detail for a study to be evaluated (Levitt et al., 2018). The proposed framework extends that

principle to the artefact level, where the evidential path of a particular generated object remains inspectable.

5.2 Structured checkpoints as technological reflexivity

The optional structured checkpoint provides the framework’s most specific extension of existing audit trails. Prompt records explain how an artefact was generated, but they do not identify which source relation changed. A compact representation can expose a transformation before fluent prose absorbs it, allowing review to focus on the relation that supports the claim.

The checkpoint does not supply independent validation because the same model may generate both prose and structure. Instead, it externalises a selected relation in a form that directs attention back to the source. Candidate records should remain provisional, while structural overlap stays separate from semantic adequacy. A simpler audit object should replace graph notation whenever it performs the same diagnostic function.

5.3 Feasibility through proportional adoption

A strong practical objection is that source-to-claim documentation could turn interpretive research into a second documentation project. The study does not establish cost-effectiveness because it did not record labour time or training demands. It does show how review can operate at different scopes. Every generated artefact received the basic source and interaction record, whereas structured comparison and reader review were applied more selectively.

Documentation should grow with a claim’s reliance on LLM output. Table 6 distinguishes three levels.

Table 6. Proportional documentation levels

Level	Suitable use	Minimum record
Exploratory	Private notes or classroom activity.	Source and prompt, followed by a dated output.
Published claim	Generated material that supports a report or article.	Exploratory record plus selected reruns and review decisions.
Reusable resource	A dataset or collaborative research object.	Published record plus version history and persistent archive.

A short decision pathway translates the proportional principle into practice, as illustrated in Table 7.

Table 7. Decision pathway for proportional documentation

Decision	Action
Does generated material influence an analytical claim?	If no, ordinary project notes are sufficient. If yes, preserve the source together with the dated interaction.
Can direct comparison with the source resolve the issue?	If yes, record the review decision and stop. If no, add a structured checkpoint.
Does the checkpoint clarify or change the decision?	If no, further documentation adds little value. If yes, retain the checkpoint with the review rationale.
Will the artefact support publication or later reuse?	If yes, add version history and a persistent archive. If no, retain the project-level record.

The pathway stops escalation when direct source comparison resolves the issue or when another record no longer clarifies the analytical decision. Structured review is reserved for cases in which a relation supporting the claim remains uncertain.

The practical claim remains limited. Selective scaling is possible, but the net cost has not been established. A future feasibility study should compare basic logging with structured review and measure both labour and changes in analytical decisions.

5.4 *Practical and ethical limits*

The diagnostic study is small. Constructed sentences cannot reproduce the density or context of an interview collection, while eight readers do not represent a research community. Their judgements illustrate how disagreement can be preserved, but broader claims require another design.

Formal comparison carries a related risk because generated graphs can repeat a shared error, while defensible representations may differ in form. Human review must remain anchored in the source (Messner et al., 2025). Access can also be constrained when source licences or participant consent restrict redistribution. In other words, the archive should record access conditions and use de-identification or access controls where needed (Autio et al., 2024).

The framework is most directly suited to textual interpretation. Note that it would need adaptation for visual or multimodal research, where recent methods centre iterative generation and researcher agency (Phan and Le, 2026; Lai and Wu, 2026).

6 Conclusion

Hosted LLMs weaken any simple link between trustworthiness and identical output, since later calls may differ while repeated prose preserves the same source distortion. Source-to-claim documentation responds by linking the research frame and model interaction to review and reanalysis.

The diagnostic demonstration showed why the links must remain visible. Stable paraphrases can drift, while quotation-compliant memos can exceed their evidence. Average graph overlap can also conceal movement among items. Structured audit identifies a relation that warrants inspection, whereas human judgement determines whether the change affects interpretation.

Humanities and social research should preserve generative mediation wherever it can alter a claim. Exploratory work may need only a light record, while publication or reuse calls for stronger documentation because auditability depends on recoverable evidence rather than identical regeneration.

AI use declaration

OpenAI models generated the research artefacts described in the diagnostic experiment. During manuscript preparation, ChatGPT provided local language suggestions and assisted with Python debugging. The logged ancillary use involved GPT-5.3 Instant and GPT-5.4 Thinking, followed by GPT-5.5 after its release. The authors reviewed every proposed change and remain responsible for the manuscript and analysis.

References

- Aguinis H, Solarino AM (2019) Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strateg Manag J* 40(8):1291–1315. <https://doi.org/10.1002/smj.3015>
- Anis S, French JA (2023) Efficient, explicatory, and equitable: Why qualitative researchers should embrace AI, but cautiously. *Bus Soc* 62(6):1139–1144. <https://doi.org/10.1177/00076503231163286>
- Autio C, Schwartz R, Dunietz J, Jain S, Stanley M, Tabassi E et al. (2024) Artificial intelligence risk management framework: Generative artificial intelligence profile. NIST AI 600-1. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.600-1>
- Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U et al. (2013) Abstract Meaning Representation for sembanking. In: *Proceedings of the 7th Linguistic Annotation*

- Workshop and Interoperability with Discourse, pp 178–186. Association for Computational Linguistics, Sofia, Bulgaria. <https://aclanthology.org/W13-2322/>
- Bender EM, Friedman B (2018) Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Trans Assoc Comput Linguist* 6:587–604. https://doi.org/10.1162/tacl_a_00041
- Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp 610–623. Association for Computing Machinery, New York, NY. <https://doi.org/10.1145/3442188.3445922>
- Berger R (2015) Now I see it, now I don’t: Researcher’s position and reflexivity in qualitative research. *Qual Res* 15(2):219–234. <https://doi.org/10.1177/1468794112468475>
- Bishop L (2007) A reflexive account of reusing qualitative data: Beyond primary/secondary dualism. *Sociol Res Online* 12(3):43–56. <https://doi.org/10.5153/sro.1553>
- Cai S, Knight K (2013) Smatch: An evaluation metric for semantic feature structures. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp 748–752. Association for Computational Linguistics, Sofia, Bulgaria. <https://aclanthology.org/P13-2131/>
- Dalkin S, Forster N, Hodgson P, Lhussier M, Carr SM (2021) Using computer assisted qualitative data analysis software (CAQDAS; NVivo) to assist in the complex process of realist theory generation, refinement and testing. *Int J Soc Res Methodol* 24(1):123–134. <https://doi.org/10.1080/13645579.2020.1803528>
- Dunivin ZO (2025) Scaling hermeneutics: A guide to qualitative coding with LLMs for reflexive content analysis. *EPJ Data Sci* 14:28. <https://doi.org/10.1140/epjds/s13688-025-00548-8>
- Fehring L, Frings J, Rust P, Kempny C, Thürmann PA, Meister S (2025) Extension of

the Consolidated Criteria for Reporting Qualitative Research guideline to large language models (COREQ+LLM): Protocol for a multiphase study. *JMIR Res Protoc* 14:e78682. <https://doi.org/10.2196/78682>

Forster E, Kartschmit N, Klager E, Mosor E, Schuster B, Mosor E, Stamm T, Donsa K (2026) Multi-stage LLM pipeline to support qualitative content analysis: A proof of concept experiment with expert validation. In: *dHealth 2026: Proceedings of the 20th Health Informatics Meets Digital Health Conference. Studies in Health Technology and Informatics*, vol 335. IOS Press, Amsterdam, the Netherlands, pp 110–116. <https://doi.org/10.3233/SHTI260065>

Geburu T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III et al. (2021) Datasheets for datasets. *Commun ACM* 64(12):86–92. <https://doi.org/10.1145/3458723>

Goodman MW (2020) Penman: An open-source library and tool for AMR graphs. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp 312–319. Association for Computational Linguistics, Online. <https://doi.org/10.18653/v1/2020.acl-demos.35>

Irwin S (2013) Qualitative secondary data analysis: Ethics, epistemology and context. *Prog Dev Stud* 13(4):295–306. <https://doi.org/10.1177/1464993413490479>

Jones KML (2025) Generative AI in qualitative research and related transparency problems: A novel heuristic for disclosing uses of AI. *Int J Qual Methods* 24. <https://doi.org/10.1177/16094069251404329>

Katz A, Fleming GC, Main JB (2026) Thematic analysis with open-source generative AI and machine learning: A new method for inductive qualitative codebook development. *Humanit Soc Sci Commun* 13:209. <https://doi.org/10.1057/s41599-026-06508-5>

Lai Z, Wu E (2026) Generative visual dialogue: An exploratory methodological framework for AI-assisted qualitative inquiry. *Int J Soc Res Methodol* 1–18. <https://doi.org/10.1080/>

- Levitt HM, Bamberg M, Creswell JW, Frost DM, Josselson R, Suárez-Orozco C (2018) Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. *Am Psychol* 73(1):26–46. <https://doi.org/10.1037/amp0000151>
- Lincoln YS, Guba EG (1985) *Naturalistic inquiry*. Sage.
- Lu MH, Ellegood R, Rodriguez-Ramirez R, Blumert S (2026) Affording process auditability with QualAnalyzer: An atomistic LLM analysis tool for qualitative research. *arXiv:2604.03820*. <https://doi.org/10.48550/arXiv.2604.03820>
- Messner R, Smith S, Richards C (2025) Artificial intelligence and qualitative data analysis: Epistemological incongruences and the future of the human experience. *Int J Qual Methods* 24. <https://doi.org/10.1177/16094069251371481>
- Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B et al. (2019) Model cards for model reporting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp 220–229. Association for Computing Machinery, New York, NY. <https://doi.org/10.1145/3287560.3287596>
- Morgan DL (2023) Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *Int J Qual Methods* 22. <https://doi.org/10.1177/16094069231211248>
- National Academies of Sciences, Engineering, and Medicine (2019) *Reproducibility and replicability in science*. The National Academies Press. <https://doi.org/10.17226/25303>
- Nowell LS, Norris JM, White DE, Moules NJ (2017) Thematic analysis: Striving to meet the trustworthiness criteria. *Int J Qual Methods* 16. <https://doi.org/10.1177/1609406917733847>
- Paulus TM, Lester JN (2024) Digital qualitative research workflows: A reflexivity framework for technological consequences. *Int J Soc Res Methodol* 27(6):621–634. <https://doi.org/10.1177/16094069241211248>

1080/13645579.2023.2237359

Phan ANQ, Le C (2026) AI as research partner: Key implications of using AI for data visualisation in qualitative research. *Int J Soc Res Methodol* 29(2):313–320. <https://doi.org/10.1080/13645579.2025.2518562>

Pratt MG, Kaplan S, Whittington R (2020) Editorial essay: The tumult over transparency: Decoupling transparency from replication in establishing trustworthy qualitative research. *Adm Sci Q* 65(1):1–19. <https://doi.org/10.1177/0001839219887663>

Rodgers BL, Cowles KV (1993) The qualitative research audit trail: A complex collection of documentation. *Res Nurs Health* 16(3):219–226. <https://doi.org/10.1002/nur.4770160309>

Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten simple rules for reproducible computational research. *PLoS Comput Biol* 9(10):e1003285. <https://doi.org/10.1371/journal.pcbi.1003285>

Schöch C (2023) Repetitive research: A conceptual space and terminology of replication, reproduction, revision, reanalysis, reinvestigation and reuse in digital humanities. *Int J Digit Humanit* 5:373–403. <https://doi.org/10.1007/s42803-023-00073-y>

Tracy SJ (2010) Qualitative quality: Eight “big-tent” criteria for excellent qualitative research. *Qual Inq* 16(10):837–851. <https://doi.org/10.1177/1077800410383121>

Van Gysel JEL, Vigus M, Chun J, Lai KH, Moeller S, Yao J et al. (2021) Designing a uniform meaning representation for natural language processing. *KI Künstl Intell* 35:343–360. <https://doi.org/10.1007/s13218-021-00722-w>

Zheng Z, Hafizuddin-Syah BAM, Tan QL (2026) Prompt-imposed structure: A methodological evaluation of LLM-generated survey data for theory-based modeling. *Int J Soc Res Methodol* 1–18. <https://doi.org/10.1080/13645579.2026.2625160>

Endnotes

¹Methodological reflexivity refers here to the deliberate examination of how research design and analytical decisions shape the resulting account. Technological reflexivity narrows that examination to the influence of digital tools and environments on those decisions and outcomes.

²A model manifest is a compact account of the model as it was available to the researcher. It records the public identifier and access date, along with the interface and observable generation controls.

³PENMAN is a parenthesized plain-text notation for storing and exchanging semantic graphs (Goodman, 2020).

⁴AMR is a rooted, directed graph for representing sentence-level concepts and their labelled relations (Banarescu et al., 2013).

⁵UMR extends graph-based meaning representation beyond the sentence by adding cross-sentence reference and other document-level relations (Van Gysel et al., 2021).

⁶Both conditions were run through OpenAI API calls rather than through the ChatGPT web interface. The model identifiers and access dates are those recorded in the API logs. Because the models were hosted by the provider, the identifiers do not denote locally preserved or immutable model snapshots.

⁷Temperature and top-p are sampling controls used by many text-generation systems. Temperature 0 reduces random variation, while top-p 1.0 leaves the full probability mass available. Neither setting guarantees identical outputs in a hosted service.

⁸A structured output is parseable when software can recover its concepts and labelled relations from the textual representation.

⁹Smatch aligns the concepts and labelled relations in two AMR-style graphs. It combines precision and recall into an F1 score between 0 and 1. The score describes structural overlap here and does not establish semantic correctness.

¹⁰The students at [institution] volunteered to participate. They served as nonexpert readers because the survey examined whether semantic drift and interpretive overreach could be recognised without graph-annotation training. The sample was diagnostic rather than representative, so claims about expert or disciplinary consensus would require a larger purposive panel.

¹¹JSON Lines, often abbreviated JSONL, is a plain-text format in which each line contains one complete JSON record.

Ethical approval

The human-participant component of this study was conducted in the United States. It was reviewed by the institutional review board responsible for the study [committee name and protocol number redacted for double-anonymized review; full details are provided in the cover letter] and was determined to be exempt from further IRB review on March 23, 2026. The exemption applied to a minimal-risk reader-judgment survey in which participants evaluated constructed textual materials and model-generated responses. The approving body was located in the United States, the same country in which the human-participant research was conducted. No human-participant research was conducted in another country, so additional local approval in multiple countries was not applicable. All procedures were conducted in accordance with applicable institutional requirements and the applicable provisions of the Federal Policy for the Protection of Human Subjects (45 CFR Part 46, Subpart A).

Informed consent

Informed consent was obtained from all participants prior to their participation in the study.