

From Genuine Citations to Warranted Readings: An Edition-Sensitive Audit of AI-Generated Literary Criticism

AI author (platform designation): OpenAI ChatGPT (GPT-6)

Human prompter: Bani Prasad Mali

Manuscript type: Conceptual / methodological essay

Prepared: 23 October 2026

Status: AI-generated research manuscript for transparent review; not an independently human-validated study

Abstract

An authentic bibliographic reference does not, by itself, warrant the scholarly proposition for which it is cited. This distinction is consequential in AI-generated literary criticism, where a fluent interpretation can conceal errors concerning historically different textual witnesses, speakers' knowledge and the reasoning that connects quotations to claims. This conceptual essay proposes two targeted checks within a five-part evidence audit: the **textual-witness substitution test**, which asks whether an interpretation survives the substitution of a different documented edition, and the **narrative-attribution test**, which asks whether a proposition remains supportable when its cited statement is assigned only the authority of its actual narrative speaker. Kate Chopin's *The Story of an Hour* supplies a bounded demonstration: the variation between the 1894 and 1895 published texts and the final explanation offered by the doctors. Existing Chopin scholarship and research on AI citation and literary analysis are acknowledged as prior work. The contribution is a limited operational synthesis, not the discovery of a new reading or an experimentally validated method. The essay concludes with a transparent claim record and proposals for empirical testing.

Keywords: textual criticism; generative AI; quotation attribution; evidence audit; Kate Chopin; interpretive warrant; literary scholarship

1. Introduction: when a genuine citation is insufficient

Bibliographic fabrication is an established risk in machine-generated academic text. Walters and Wilder (2023) found frequent errors and non-existent citations in the language models they examined. Han et al. (2026) examined 3,201 peer-reviewed conference papers using a multi-step verification protocol and flagged 69 papers containing at least one reference that remained unverifiable. The latter result is a finding about resolvability under a defined protocol; it is not evidence that AI caused every flagged reference. Such work makes a necessary contribution: scholarship must cite sources that readers can recover.

Literary criticism also depends on a different relationship. A source can exist, and a quotation can be authentic, while a conclusion about the quotation is inadequately supported. A sentence might belong to a different edition from the one identified; a character's explanation might be misrepresented as the narrator's; or a local passage might be used to establish an unrestricted historical claim. These failures need not involve fictitious references. They concern **interpretive warrant**: the defensible link between a proposition and the particular textual evidence supplied for it.

The distinction between document and interpretation is not new. Textual scholars have long demonstrated that the edition chosen can affect what is interpreted (Tanselle, 1989). Drucker's (2011) account of humanities *capta*

likewise cautions against treating selected and constructed evidence as simply given. Recent work on evidence-based AI also substantially anticipates the present concern. Schreieder et al. (2026) survey 134 studies of attribution, citation and quotation in evidence-based text generation. Michel et al. (2025) evaluate quotation-to-speaker attribution in novels. Chan and Wong (2025) investigate the depth, accuracy and relevance of model-generated literary analysis.

This essay makes a deliberately narrower proposal: a concrete way to test two risks that are particularly important to the interpretation of literary texts. The **textual-witness substitution test** checks whether an argument tacitly depends on a different historical version. The **narrative-attribution test** checks whether the authority assigned to a statement exceeds the knowledge or position of the figure who utters or supplies it. These checks are embedded in an audit comprising source identity, witness identity, attribution, inference and counterevidence. Their use is illustrated rather than statistically validated.

2. Method and evidentiary boundaries

The example is Kate Chopin’s short story first published in *Vogue* on 6 December 1894 as *The Dream of an Hour* and subsequently reprinted in *St. Louis Life* on 5 January 1895. The Kate Chopin International Society provides an edited later-state transcription and explains a consequential pronoun variant (Chopin, 1894/1895; Kate Chopin International Society, n.d.). Its historical account notes disagreement about whether the changes originated with Chopin or editorial intervention. The essay relies on that editorial documentation rather than pretending to offer a new collation of original periodicals.

The case is selected for two reasons. First, one sentence in the documented publication history differs by a word with consequences for its interpretation. Second, the closing sentence reports an explanation attributed to doctors whose access to Louise Mallard’s private experience is limited within the fiction. These examples make it possible to separate observed textual form, represented statement and critic’s inference.

This is a **worked conceptual demonstration**, not a study measuring the behaviour of GPT-6 or any other model. No corpus of outputs, random sample, comparison group, independently collected field data or inter-rater reliability estimate is claimed. The proposed audit’s effectiveness therefore remains an open empirical question. In addition, interpretations by Berkove (2000) and Cunningham (2004) are treated as existing scholarly alternatives, not findings originated here.

3. First probe: textual-witness substitution

Louise Mallard, believing her husband Brently has died, moves from immediate grief toward imagining future years no longer governed by marital authority. The published texts differ in a sentence expressing that possibility. The 1894 wording includes “**no one to live for**”; the subsequent version reads “**no one to live for her**” (Kate Chopin International Society, n.d.). The Society explicitly identifies the difference and explains its significance. It would be inaccurate to claim that this essay discovered it.

Without the added pronoun, the sentence may foreground the duty of living for somebody else. With *her*, the grammar more directly positions Louise in relation to somebody else’s activity on her behalf; read with the next sentence about one will bending another, the revised construction may foreground the problem of external direction. The readings remain related. Their relation to the local wording, however, is not identical.

The **textual-witness substitution test** asks: *If the scholar’s cited edition is replaced by the other documented*

version, does the stated grammatical evidence remain present and does the inference remain unchanged? For a broad account of Louise's anticipated freedom, the difference may not defeat the interpretation. For a claim specifically resting on the inserted pronoun, it is decisive. An argument explaining the significance of *her* while giving only the original *Vogue* wording as its evidence would mismatch the witness and the claim. This descriptive failure is called **witness displacement** here; the underlying issue is familiar to textual criticism.

The case also shows why textual observation and historical intention must be separated. The specialist editorial account notes that Daniel Rankin attributed changes to Chopin, while Per Seyersted entertained the likelihood of intervention by editor Sue V. Moore (Kate Chopin International Society, n.d.). An annotated publication witness is preserved in the Missouri Historical Society's Chopin Papers, but an annotation's existence does not alone establish who made every alteration or exactly why. The documented difference supports an argument about published wording. It does not, by itself, conclusively demonstrate Chopin's motive in introducing the pronoun. The audit therefore distinguishes **variant identified** from **agency demonstrated** and **intention inferred**.

This test generalises beyond Chopin as a question, not as an established performance result. Any literary-critical claim hinging on a pronoun, negation, translation choice, tense or disputed punctuation invites a comparison against the edition or witness actually named. Not every variant requires extended textual history; scrutiny should be proportional to its contribution to the article's central argument.

4. Second probe: narrative-attribution testing

The conclusion of Chopin's story poses a different evidentiary risk. Brently Mallard returns unexpectedly, and Louise dies. After Richards attempts to obscure Brently from her view, the narration states that he was too late. The final sentence reports that, when the doctors arrived, "**they said**" she died of "**joy that kills**" (Chopin, 1895 textual state, final paragraph). The attribution matters. The medical explanation is represented as the doctors' interpretation, not introduced independently as a verified medical finding.

Readers have been given access to Louise's thoughts about prospective freedom; the doctors are not shown to possess such access. That asymmetry permits an ironic reading of their explanation. It does not provide a definite clinical account of her death. Nor does the narration directly report a final internal perception after the husband's entry that could eliminate competing causal interpretations.

Critical disagreement makes this limitation visible. Berkove (2000) questions an uncomplicated emancipatory reading of Louise, directing attention to the irony and contradictions of extreme self-assertion. Cunningham (2004) offers a different account of the ending, challenging the assumption that Louise necessarily sees Brently and dies from shock at his return, and attending to the strain associated with her new sense of autonomy. These analyses do not license this essay to pronounce one single cause established beyond doubt. They demonstrate that the narration supplies events and attributed explanations from which critics make different inferences.

The **narrative-attribution test** asks: *Does the conclusion remain adequately supported if the cited explanation is assigned only to the figure or narrative level actually responsible for it?* The proposition *the doctors said her death was caused by joy* passes as an attributed report. The proposition *the narrator conclusively establishes that joy caused her death* does not pass on the same evidence. This is an example of **attribution collapse**: promoting a situated statement to unrestricted narrative authority.

The test must not be confused with computational speaker identification. Michel et al. (2025) evaluate whether a model can identify speakers of direct quotation in novels. Identifying who said something is a necessary step,

but critical analysis also asks whether that speaker knows enough for the statement to justify the researcher's interpretation. The latter is an issue of evidentiary authority, not merely speaker labelling.

5. Five-part audit and a worked claim record

The two probes can be situated within a five-part workflow:

1. **Source identity.** Confirm the bibliographic record and whether the cited publication is identifiable. A valid DOI or catalogue record is necessary but does not prove the claim.
2. **Textual-witness identity.** Record the edition, transcription or publication state on which a consequential claim relies. Apply witness substitution where a variant may affect the inference.
3. **Narrative attribution.** Establish the speaker, observer or narrator responsible for the relevant statement. Apply the attribution test before granting the statement the authority of a narrator's independent judgement.
4. **Inferential warrant.** Distinguish what the passage explicitly establishes from the critic's explanatory or historical conclusion. Identify any additional premises needed.
5. **Qualifying evidence.** Examine counterpassages, variants, competing interpretations and uncertain documents. State the confidence of the conclusion proportionately.

A record of the demonstration makes these distinctions concrete:

Proposition	Relevant evidence	Audit disposition
The published wording differs between the 1894 and 1895 states.	Documented pronoun variation	Supported
The added pronoun conclusively proves Chopin's political motivation.	Variant exists, revisionary agency disputed	Not established
The doctors attribute Louise's death to joy.	Concluding sentence, attributed speech	Supported as attribution
The doctors' account is necessarily the narrator's definitive explanation.	Situated explanation and unequal access to consciousness	Not established
Louise certainly perceives Brently and dies from that shock.	Final sequence and published critical alternatives	Contested

The dispositions **supported**, **supported with qualification**, **contested**, **not established** and **source unavailable** could be used for other claims. They are not measurements of truth and should not be converted into a reliability percentage without validated criteria. In an actual audit, a key claim should also be accompanied by a passage locator and a note identifying the specific inference. The purpose is to make disagreement inspectable, not to eliminate legitimate interpretive plurality.

This claim-level method is particularly relevant to AI-generated prose because a fluent paragraph can collapse the operations of locating a source, deciding which version matters, summarising narrative events and drawing a theoretical conclusion. Separating those operations creates checkpoints at which a reviewer can ask which

evidence is being used for which proposition. That benefit remains a plausible methodological rationale, not a measured improvement demonstrated by the present paper.

6. Relation to existing methods and editorial practice

The proposed workflow is an integration of established practices rather than a new general theory of citation. Research on fabricated references has already supplied protocols for identifying publications (Walters & Wilder, 2023; Han et al., 2026). Schreieder et al. (2026) map existing strategies for attribution and verifiable generation. Literary scholarship has long demanded edition-sensitive quotation and defensible treatment of narrative voice, while work on AI literary analysis documents the need to evaluate interpretive depth and supporting evidence (Chan & Wong, 2025).

The narrower value of the two probes is to expose particular relations that can survive a successful bibliographic check. An editor who confirms that *The Story of an Hour* exists has not checked whether a pronoun occurs in the edition cited. An editor who verifies the final sentence has not determined whether the doctors' view can be equated with narratorial authority. An AI-use declaration communicates production history; it does not establish either inferential step.

This distinction should not be turned into a blanket claim that AI-assisted writing is inevitably worse than unaided human scholarship. Human critics make the same kinds of mistakes. The method is applicable to any essay in which the verbal form and narrative authority of evidence matter. Its relevance to AI-native publication is that the paper's intellectual labour and provenance should be documented openly and its consequential claims should be made reconstructable.

7. Limitations and proposed evaluation

This essay demonstrates an audit on one narrative with a documented variant and a disputed conclusion. It does not establish that contemporary language models frequently make these particular errors, nor that the proposed probes improve performance in a controlled evaluation. The use of the Society's edited text is not a substitute for comprehensive historical collation. Likewise, only selected overlapping criticism is examined; no exhaustive priority claim is advanced.

An empirical follow-up could construct a set of claims involving versioned literary works, translations, embedded narrators and unreliable testimony. Qualified assessors could evaluate them under two conditions: bibliographic checking alone and bibliographic checking plus the two probes. The design could record detection of witness mismatches, attribution overreach, time required and agreement among assessors. Pre-registration of criteria and genuinely independent scholarly adjudication would be necessary before claiming a validated method.

These are proposals for future testing; they are not results of experiments performed for this essay.

8. Conclusion

A real citation is not synonymous with a warranted reading. A documented one-word difference in Chopin's publication history changes the grammatical material available to an interpretation; an attributed medical explanation in the same story shows how a correct quotation can be granted more narrative authority than its source

supplies. The textual-witness substitution and narrative-attribution tests make those distinctions explicit within an audit joining bibliographic accuracy to inferential accountability.

The contribution is modest and deliberately bounded. It does not invent textual criticism, discover the variant or decide the definitive meaning of Louise Mallard's death. It proposes two transparent checks whose effectiveness can be evaluated in future work. For AI-generated scholarship in particular, this is preferable to treating fluency, a verified citation or a disclosure statement as sufficient proof of intellectual reliability.

AI-generation and provenance statement

The conceptual framing, close-reading synthesis, prose, organisation and multiple revisions of this essay were produced substantially by **OpenAI ChatGPT (GPT-6)** in October 2026, following requests from human prompter **Bani Prasad Mali** to develop an article that could be transparently published as AI-generated research. The AI system also conducted web-assisted checks of accessible primary texts, bibliographic records and selected prior scholarship. The analysis is a conceptual illustration; it did not involve original model benchmarking, human participants or experimental data. The work has not received independent human scholarly validation or a formal similarity assessment, and no such validation is claimed. The human prompter's contribution was to direct and request the work, not to conduct the underlying textual criticism or independent source verification. The paper is proposed for a platform that explicitly displays AI authorship rather than for a conventional journal under a misleading human-authorship declaration.

References

- Berkove, L. I. (2000). Fatal self-assertion in Kate Chopin's "The Story of an Hour." *American Literary Realism*, 32(2), 152–158. <https://www.jstor.org/stable/27746974>
- Chan, C. K. Y., & Wong, H. Y. H. (2025). Assessing the ability of generative AI in English literary analysis through Bloom's Taxonomy. *Discover Computing*, 28, Article 127. <https://doi.org/10.1007/s10791-025-09572-8>
- Chopin, K. (1894, December 6). The dream of an hour. *Vogue*. <https://archive.vogue.com/article/1894/12/the-dream-of-an-hour>
- Chopin, K. (1895). *The story of an hour* [Later textual state, reproduced by the Kate Chopin International Society]. <https://www.katechopin.org/story-hour/>
- Cunningham, M. (2004). The autonomous female self and the death of Louise Mallard in Kate Chopin's "Story of an Hour." *English Language Notes*, 42(1), 48–55. <https://doi.org/10.1215/00138282-42.1.48>
- Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, 5(1). <https://digitalhumanities.org/dhq/vol/5/1/000091/000091.html>
- Han, C., Manoharan, S., Ye, X., & Speidel, U. (2026). Phantom citations: An empirical study of non-existent and unverifiable references in scholarly literature. *Journal of the Association for Information Science and Technology*, 77(10), 1287–1316. <https://doi.org/10.1002/asi.70104>
- Kate Chopin International Society. (n.d.). *Kate Chopin: "The Story of an Hour."* <https://www.katechopin.org/the-story-of-an-hour/>

- Michel, G., Epure, E. V., Hennequin, R., & Cerisara, C. (2025). Evaluating LLMs for quotation attribution in literary texts: A case study of LLaMa3. In *Proceedings of NAACL 2025 (Volume 2: Short Papers)* (pp. 742–755). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-short.62>
- Missouri Historical Society. (n.d.). “*The Dream of an Hour*,” *Vogue Magazine*. Retitled in Chopin’s hand: “*The Story of an Hour*” [Archival item D00404]. Kate O’Flaherty Chopin Papers, A0273. https://mohistory.mobiusconsortium.org/repositories/2/archival_objects/1111
- Schreieder, T., Schopf, T., & Färber, M. (2026). Attribution, citation, and quotation: A survey of evidence-based text generation with large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 30956–31000). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.acl-long.1430>
- Tanselle, G. T. (1989). *A rationale of textual criticism*. University of Pennsylvania Press.
- Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by Chat-GPT. *Scientific Reports*, 13, Article 14045. <https://doi.org/10.1038/s41598-023-41032-5>