What Happens to Citations When AI Rewrites Academic Text? A Large-Scale Paired Audit

TextPulse Research · Working Paper · doi:10.5281/zenodo.22017214

Abstract

Published audits of AI-generated references examine how often language models invent citations when writing new text. This study examines a different and more common situation. When a language model rewrites academic text that already contains citations, do those citations change? We audited 60,786 paired academic research snippets. Each pair consists of a human-written academic excerpt and a machine rewrite of that same excerpt, generated by eight model configurations across six model families. The pairs contain 213,881 in-text citation marks, and every citation in a rewrite can be checked exactly against the source excerpt it came from. In total, 96.9 percent of citations survived rewriting unchanged and a further 0.8 percent survived as format conversions. The remaining 2.26 percent (95 percent confidence interval 2.19 to 2.32) were corrupted. In total, 1.30 percent of citations were dropped, 0.53 percent appeared in rewrites with no counterpart in the source (fabricated), and 0.42 percent were altered inline, including corrupted author names and changed years. Corruption rates varied drastically by model, from 0.30 percent for the most faithful configuration to 8.34 percent for the least, a 28 times difference on the same inputs. We also identify a small class of rewrites that replace an anonymous numeric citation with a specific author-year claim that the source text never stated. All verdicts are released publicly as a per-citation evidence file.

1. Introduction

Most real-world use of language models on academic text is not generation, but transformation. Writers paste in existing drafts, paragraphs, and methods sections and ask a model to refine, shorten, expand, formalize, or paraphrase them. This text carries citations, and the citations attribute claims to real publications and authors. A citation that does not survive the transformation damages the academic integrity of the document as surely as a fabricated one, and the damage is the same problem now being measured at literature scale, where fabricated and unverifiable citations are entering published biomedical papers at an alarmingly rising rate (Topaz et al., 2026).

The literature on hallucinated references has established that models requested to generate references from scratch invent a significant share of them. However, that literature does not cover the transformation case. When a model rewrites a passage that contains a citation, the citation is present in the context window and only needs to be copied. Investigation on whether models reliably copy it has not yet been measured, and the answer matters to a large population of

academic users, including anyone who processes cited text through a paraphrasing tool, an AI editing platform, or an AI chat model (e.g., GPT, Claude or Gemini).

This study addresses four main questions. First, how often do citations survive an AI rewrite intact? Second, when they do not survive, what happens to them? Third, how much does fidelity differ across models on the same inputs? Fourth, do citation formats differ in how well they survive?

Since the source text is available for every AI rewrite, verification does not require an external database. The ground truth for each rewrite citation is the original text snippet it was rewritten from. This allows an audit at a scale of over 200,000 citations, which generation-side studies cannot reach because each of their references must be checked against bibliographic databases, a tedious effort.

2. Background on citation formats

The outcome this study measures is defined by citation conventions, and so we summarize the formats involved. An in-text citation is a compact pointer. It identifies one entry in the document's reference list, and the entry carries the full bibliographic record, including the DOI. The citation can therefore be considered a key. If a rewrite corrupts the key, the reference list entry is still present and still correct, but nothing in the text points to it anymore, and the claim it supported becomes entirely unattributed. This is the reason citations deserve their own fidelity audit, separate from any question about whether the corresponding references are real.

Academic content uses three broad families of citation formats, and all three appear in this study's corpus. Author-year styles such as APA render citations either parenthetically, with authors and year within parentheses at the end of a clause, as in "(Moffitt & Tormey, 2014)", or narratively, with the authors acting as part of the sentence and only the year in parentheses, as in "Moffitt and Tormey (2014) argue". Numeric styles replace names with an index into a numbered reference list, in brackets. Superscript styles do the same with raised indices attached to the end of a sentence. Each format carries different amounts of information in the citation itself, which turns out to matter for how well it survives rewriting.

Author-year citations also carry systematic surface variety that a rewriting model must navigate. APA's 6th edition listed up to five authors in the first mention, while the 7th edition shortens three or more authors to "et al." from the first citation, uses an ampersand inside parentheses but "and" in narrative form, and attaches possessives to narrative citations, as in "Hofstede's (1984) framework" (American Psychological Association, 2010, 2020). A faithful rewrite may change between these formats, while preserving the same corresponding reference. The audit's verdict scheme is built around this distinction. Format changes that keep the reference intact are counted separately from changes that corrupt or lose it.

Numeric and superscript citations are the extreme case of compression. The citation carries no bibliographic information at all, only a numbered position in a list. A model that changes a numeric index, or re-renders a superscript as a bracketed number of its own choosing, repoints the sentence at a different reference. A model that replaces a numeric mark with an author-year claim is not converting a format. It is asserting bibliographic detail that the passage never contained in the first place, a behavior this study investigates and reports individually.

3. Literature review

The published work on citations and language models is almost entirely based on AI generation. Audits of references produced from scratch found high fabrication rates in 2023 models. Walters and Wilder (2023) verified 636 citations from generated literature reviews and found 55 percent of GPT-3.5 citations and 18 percent of GPT-4 citations fabricated. Medical content audits found comparable or higher rates, namely, 47 percent fabricated in Bhattacharyya et al. (2023) and 69 percent in Gravel et al. (2023). Subsequent work extended these findings across AI model vendors and academic disciplines. A cross-AI model audit of 400 references from eight systems found only 26.5 percent fully correct (Cabezas-Clavijo and Sidorenko-Bautista, 2025), the GhostCite benchmark measured hallucination rates from 14.2 to 94.9 percent across 375,440 citations generated by 13 models (Xu et al., 2026), and retrieval-based legal research tools still hallucinate in 17 to 33 percent of responses (Magesh et al., 2025). Our companion study (TextPulse Research, 2026b) measures current-generation fabrication under a single cross-family protocol and finds that fabrication currently concentrates where full bibliographic detail is requested. On the receiving end, fabricated citations are entering the published record itself (Topaz et al., 2026). Surveys of hallucination provide the general framework (Ji et al., 2023).

However, the transformation case, where an AI model is requested to rewrite cited text, has not yet been measured. Paraphrasing research has concentrated on whether rewritten text evades AI detectors (Krishna et al., 2023) and on the general semantic fidelity of rewriting systems, but not on the survival of in-text citations. Studies of citation preservation or corruption under machine paraphrase are lacking. This study fills that gap with a paired design in which every rewrite is checked against its original source passage, which removes the need for external verification of the primary outcome and allows for an audit at a scale two orders of magnitude beyond generation-side studies.

4. Data and methods

4.1 Corpus

The audit uses an internal corpus of 60,786 human-AI paraphrase pairs built from open-access academic text. Each pair comprises a source passage of roughly 100 to 400 words extracted from published scholarly writing in STEM, medicine, social science, economics, and humanities

disciplines, and a rewrite of that passage generated by one of eight model configurations. Corpus 1 contains 38,323 pairs produced by two configurations, a DeepSeek chat model and a Gemini Flash-class model. Corpus 2 contains 22,463 pairs produced by six configurations from the DeepSeek, Grok, Mistral, Qwen, Gemini, and OpenAI families. Rewrites were generated with simple one-shot instruction prompts that asked the model to rephrase the passage while preserving its content. The two corpora used different prompt regimes, so model comparisons are reported within each corpus. Corpus 2, in which all six families rewrote under the same regime, carries the cross-model comparison.

Of the 60,786 pairs, 34,211 contain at least one in-text citation, contributing 213,881 citations to the audit. Source passages carry three citation formats, which are parenthetical author-year citations, narrative citations, and numeric citations. The latter mostly appear both as bracketed marks and as superscript indices flattened to trailing digits during text extraction.

4.2 Extraction and alignment

A rule-based extractor parses citation marks from both sides of every pair. It covers parenthetical author-year items, including multi-citation parentheticals and items preceded by connecting phrases, narrative citations including possessive forms, bracketed numeric marks, and superscript-style indices with a guard against figure and table labels.

Each rewrite citation is aligned to source citations in three passes. Pass one matches within format class. It applies exact author-year matches first, then fuzzy author matches at the same year with a normalized edit similarity of at least 0.72, then same-author matches at a different year, and for numeric marks exact and overlapping index sets. Pass two pairs the remaining unmatched citations across format classes as style conversions. Pass three is a verbatim-presence fallback. Before any rewrite citation is classified as invented, the source text is searched for the citation's lead author. If the author is present, the citation is classified as preserved, or as a year mutation if the claimed year appears nowhere in the source. The fallback runs before a source citation is classified as dropped.

Every mark receives one verdict. The verdicts are PRESERVED, MUTATED-AUTHOR, MUTATED-YEAR, MUTATED-NUMERIC, CONVERTED-STYLE for the same reference re-rendered in a different citation format, CONVERTED-ATTRIBUTION for a numeric source mark re-rendered as a specific author-year claim, INVENTED for a rewrite citation with no counterpart in the source, and DROPPED for a source citation with no counterpart in the rewrite.

4.3 Classifier validation

The classifier was validated over the full population of non-preserved verdicts rather than a sample. For each of the INVENTED verdicts, we tested if the citation's lead author appeared anywhere in the source text. For each DROPPED verdict, we tested if the source citation's lead author appears anywhere in the rewrite. Any hit marks the verdict as a suspect. After the fallback pass, residual

suspect rates were 5.4 percent for author-year INVENTED verdicts (3 of 56) and zero for narrative INVENTED verdicts. Manual inspection demonstrated that the residual suspects are mostly genuine near-miss mutations that string search detects but strict alignment does not. Examples include a rewrite citing “Karin-D’Arcu, 2005” against the original correct source containing “Karin-D’Arcy, 2005”. The corruption rates reported below are thus undercounts, and not overcounts. A prior version of the pipeline without the fallback pass overcounted inventions by over five times. The validation procedure and both sets of counts are included in the released code.

4.4 Statistics

Rates are reported with Wilson 95 percent confidence intervals. The corrupted-mark rate counts the INVENTED, DROPPED, and three MUTATED classes. CONVERTED-STYLE is reported separately as a formatting change. CONVERTED-ATTRIBUTION is reported separately as an unverifiable attribution. Model differences within corpus 2 are tested with a chi-square test on preserved vs. corrupted counts.

5. Results

5.1 Overall fidelity

Of 213,881 citations, 207,229 (96.9 percent) were preserved intact and 1,771 (0.8 percent) were preserved as format conversions. The corrupted remainder amounts to 2.26 percent of all citations (95 percent confidence interval 2.19 to 2.32). In it, 2,776 source citations were dropped (1.30 percent), 1,141 rewrite citations had no source counterpart (0.53 percent), and 907 marks were altered in place (0.42 percent), of which 553 corrupted the author string, 188 shifted the year, and 166 corrupted a numeric index set. A further 57 marks were converted attributions, which are described in section 5.4.

5.2 Differences between models

Fidelity depends on the model used to generate the text, not on the text itself. On identical source texts within corpus 2, corrupted-citation rates ranged from 0.30 percent for Grok (confidence interval 0.25 to 0.37) to 8.34 percent for Qwen (7.85 to 8.86), a 28-fold difference (chi-square 2506.4, df 5, n 117,147). The full ordering was Grok 0.30, Gemini 0.93, DeepSeek 3.99, OpenAI 5.34, Mistral 6.67, and Qwen 8.34 percent. Corpus 1, generated under a prompt designed for faithful rewriting, shows that low corruption is achievable at scale. Its two configurations produced 0.35 and 0.41 percent corrupted citations across almost 97,000.

The composition of corruption also differs by model. The Qwen and OpenAI configurations dropped citations at the highest rates, namely, 5.7 and 4.9 percent of their citations. Mistral combined frequent dropping with the highest share of citations that lack a source counterpart. Grok corrupted almost nothing in any category.

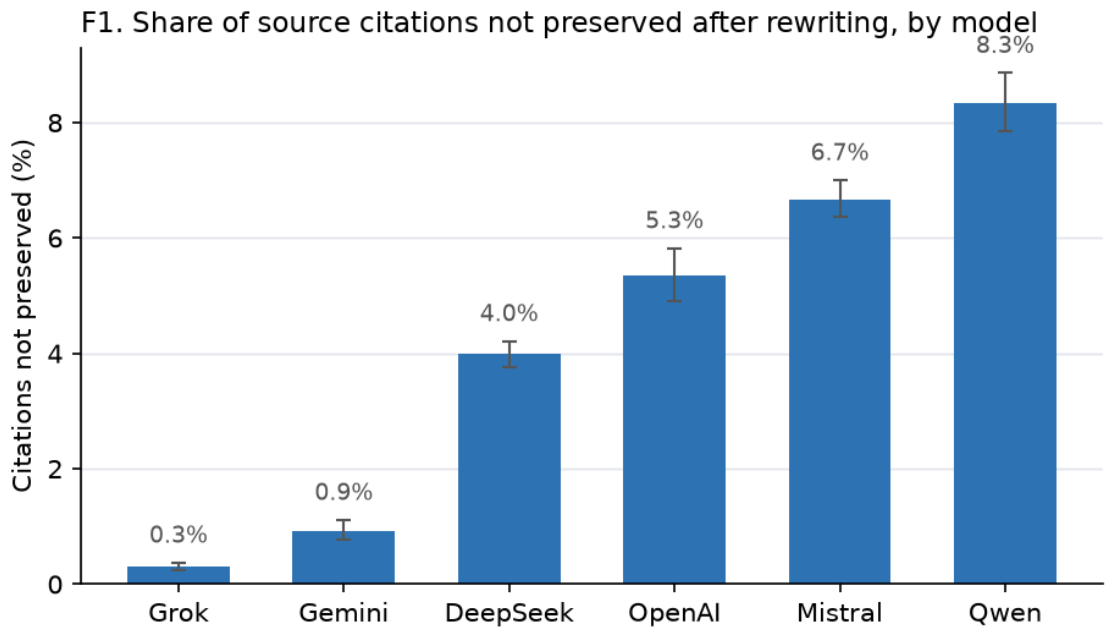

Figure F1. Share of source citations not preserved after rewriting, by model, with 95 percent confidence intervals.

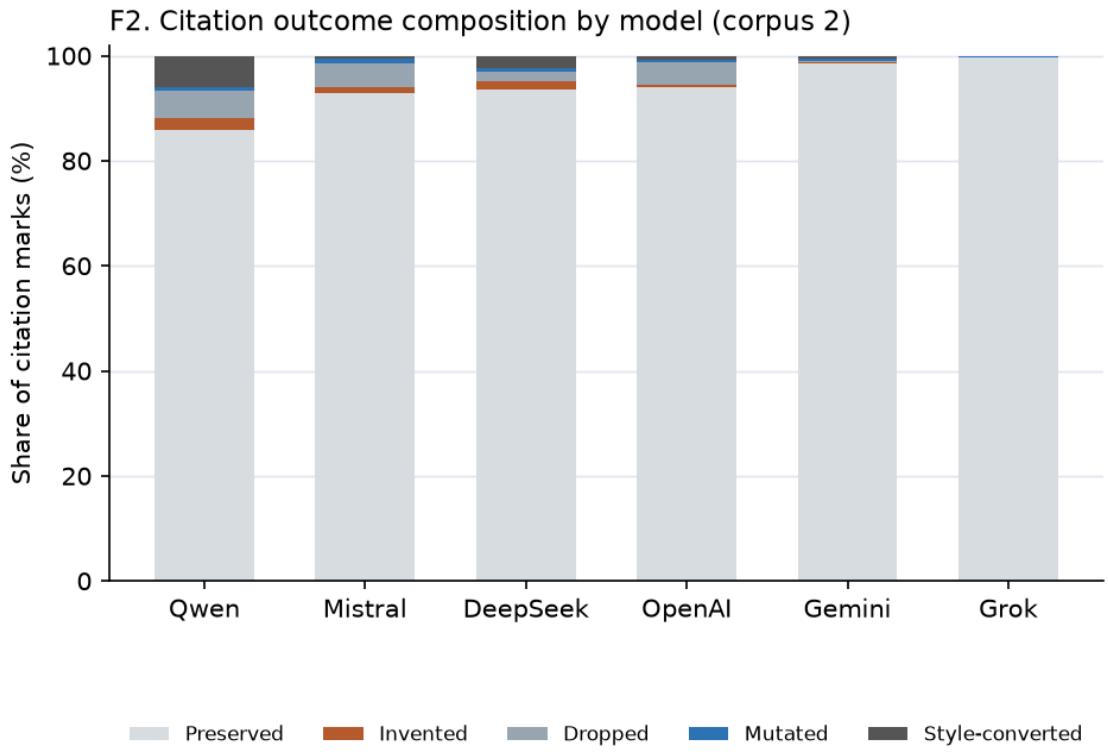

Figure F2. Citation outcome composition by model in corpus 2.

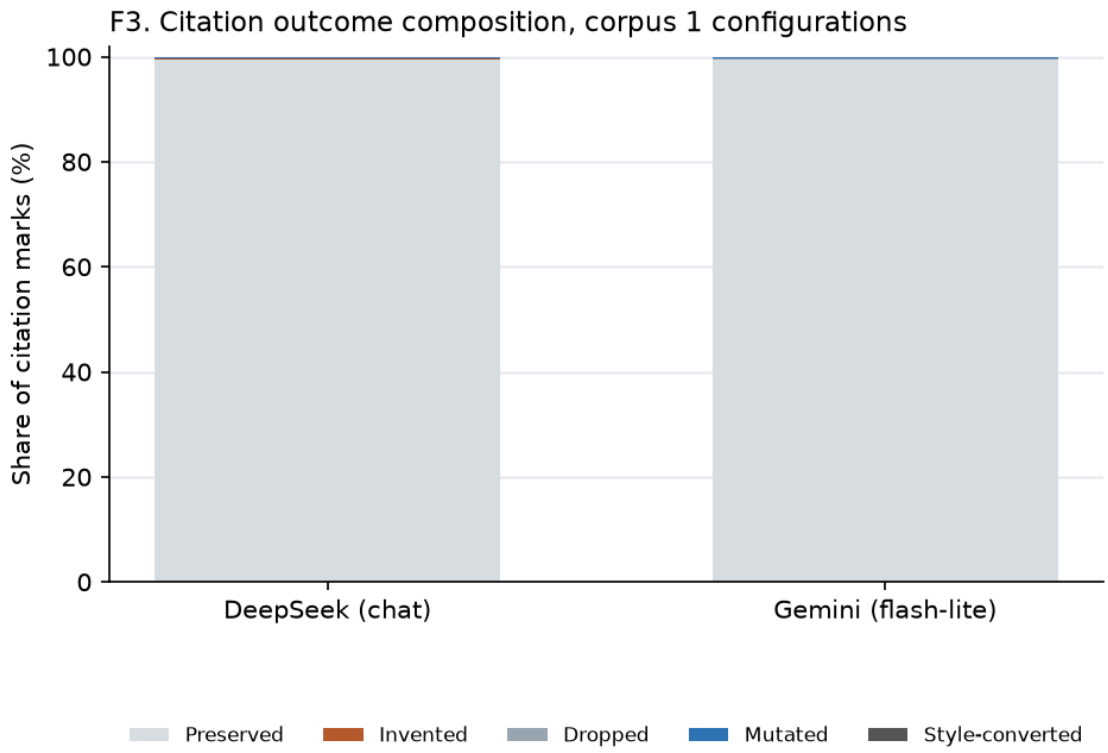

Figure F3. Citation outcome composition for the two corpus 1 configurations.

5.3 Author and year mutations

Author mutations are correct-looking spelling changes rather than random noise. The audit found doubled or simplified letters, stripped or garbled diacritics, apostrophe variants, and in multi-author strings, one surname assimilated to a neighboring one. Figure F5 shows a representative case from the evidence file. A humanities text citing “(Moffitt and Tormey 2014)” came back from the Gemini configuration citing “(Moffitt and Tormitt 2014)”, the second surname reshaped toward the first. Other cases in the released data include “Gupta et al. 2009” rewritten as “Gupert et al. 2009” and a source list ending “Laine, and Lecthinen” rewritten as “Laine, and Laine”. These mutations are the most consequential corruption class because the citation still appears to be complete. The error only appears when a reader checks the reference list, and the lookup then fails without explanation. Year mutations behave similarly and typically change a citation by one to a few years.

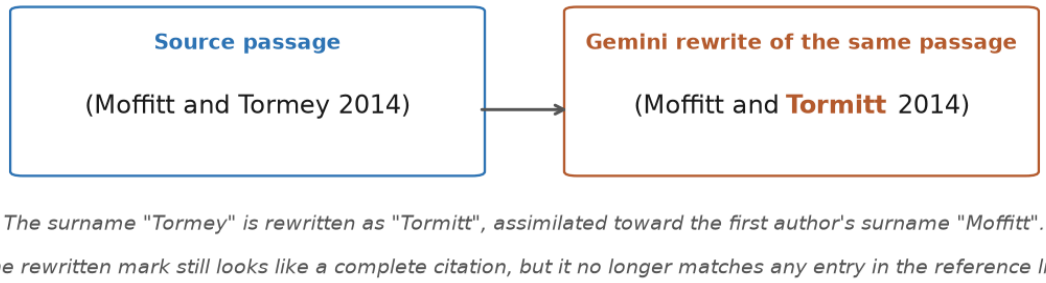

Figure F5. A source citation and its corrupted rewrite, from the released evidence file.

Some author changes were changed in the opposite direction. When the source text itself misspells a name, some models normalize it to the canonical spelling, such as “Miles and Hubermen” rewritten as “Miles and Huberman”. Under this audit’s source-fidelity criterion, these changes count as mutations, since the rewrite no longer matches the text it was given, although the change corrects the bibliographic record. Cases of this type are a small fraction of author mutations. A corrected name no longer matches the document’s own reference list entry, so even a helpful change can ‘orphan’ a citation.

5.4 Converted attribution

In a total of 57 cases, a rewrite rendered an anonymous numeric source citation as a specific author-year citation. For instance, a superscript index became “(Jain et al., 2017)”. The attribution may be correct if the model has seen the cited work, but nothing in the source text supports it. The model claims bibliographic detail from its training data inside a task presented as rewriting. This is generation-style hallucination risk appearing inside a transformation task, and it cannot be detected by comparing citation counts.

5.5 Differences between citation formats

Citation formats differ in how well they survive. In corpus 1, where the source citation style is tagged, citations in texts with numeric citations were corrupted at 0.5 percent compared with 0.3 percent for author-year texts. Superscript indices were the weakest format. Flattened to bare digits, they were the class that models most often dropped entirely or re-rendered as bracketed marks with new index numbers. Each re-rendering breaks the link to the original reference list, unless the numbering is reconstructed exactly.

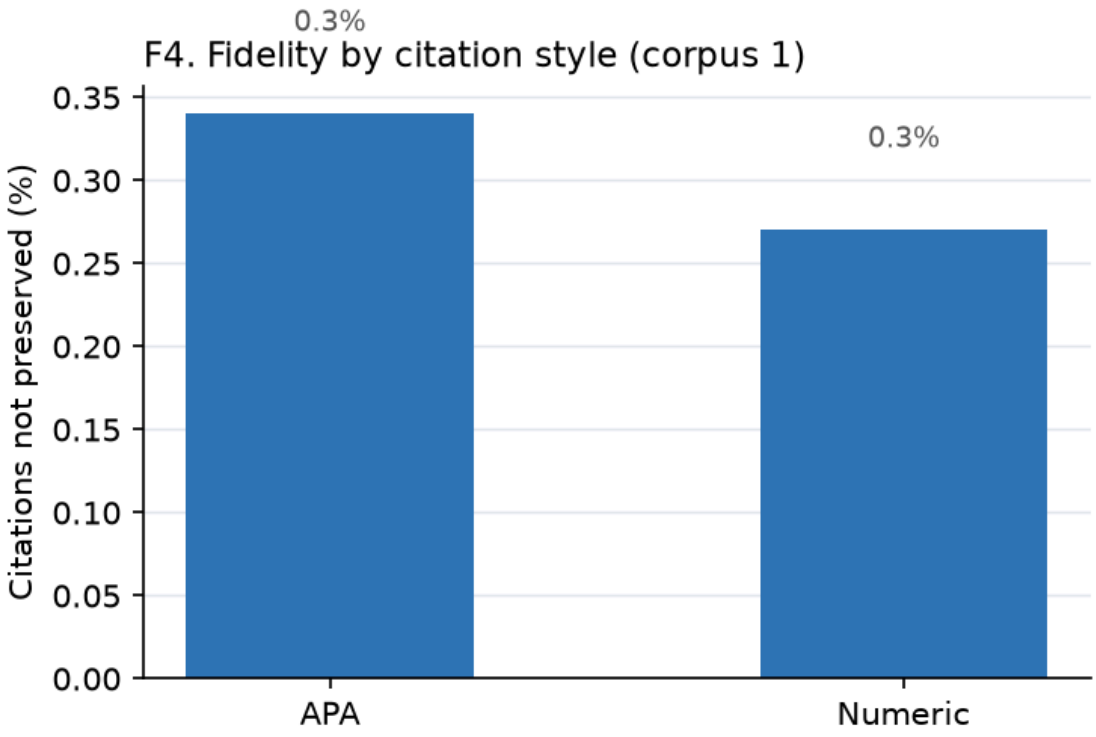

Figure F4. Corruption rate by tagged citation style in corpus 1.

6. Discussion

The results obtained highlight three primary findings. First, the overall rate is low, but the volume is not. A corruption rate of 2.26 percent means one damaged citation for roughly every 44 citations. A literature review with 80 citations, run through the least faithful configuration in this audit, would return with six or seven citations dropped, damaged, or unsupported.

Second, the difference between models is the actionable result. On identical inputs, the most faithful configuration corrupted 1 mark in 330 and the least faithful 1 in 12. Citation fidelity is an engineering property that varies by more than an order of magnitude across tools that advertise the same capability, and nothing is indicated to the user on which regime an AI model operates in. The corpus 1 results show that corruption below 0.5 percent is achievable at scale when rewriting is explicitly constrained toward fidelity.

Third, the corruption is hard to detect. Dropped citations leave prose that reads as unattributed common knowledge. Mutated author names look intact and fail only on lookup. Converted attributions read as more precise than the source while asserting details the source never contained. None of these failures is visible upon a brief read of the text, which is the reason transformation-based fidelity (i.e., rewriting from a source) requires separate measurement from generation-side hallucination (i.e., generation from scratch). It is also the reason these failures can be expected to propagate. A mutated citation that survives drafting also survives review for the

same reason the author missed it in the first place, and audits of the published literature already find unverifiable citations accumulating there at an increasing rate (Topaz et al., 2026).

This study has limitations. The two corpora used different rewrite prompt regimes, so models are compared only within each corpus individually. The extractor targets the citation formats of English-language academic content. Residual classifier error, limited by the full-population validation, means the reported rates undercount true corruption. The audit measures fidelity to the source text, and not the correctness of the source's own citations. Finally, rewrite quality depends on prompt wording, so the reported rates characterize the audited configurations under their exact prompts rather than any model under all prompts.

7. Conclusion

Citations generally survive AI rewriting. The failures are rare enough to be overlooked by the user, which is what makes them consequential. Across 213,881 citation marks, one in 44 did not survive, the failure rate varied 28-fold across models on the same text, and a small class of rewrites added attributions their source never stated. Therefore, practically, citations should be checked after any AI rewrite, and tools that transform academic text should treat citation fidelity as a measured property rather than an assumed one.

Data availability

The per-citation evidence file, with one row per citation mark containing an opaque pair identifier, corpus, model, discipline, citation format, verdict, and match evidence, is openly available on Zenodo together with the extraction, alignment, validation, analysis, and figure code (https://doi.org/10.5281/zenodo.22017214). The source and rewrite passage texts are not released.

References

American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.). American Psychological Association.

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). American Psychological Association. https://doi.org/10.1037/0000165-000

Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus, 15(5), e39238. https://doi.org/10.7759/cureus.39238

Cabezas-Clavijo, Á., & Sidorenko-Bautista, P. (2025). Assessing the performance of 8 AI chatbots in bibliographic reference retrieval: Grok and DeepSeek outperform ChatGPT, but none are fully accurate. arXiv:2505.18059.

Gravel, J., D'Amours-Gravel, M., & Osmanlliu, E. (2023). Learning to fake it: Limited responses and fabricated references provided by ChatGPT for medical questions. Mayo Clinic Proceedings: Digital Health, 1(3), 226-234. https://doi.org/10.1016/j.mcpdig.2023.05.004

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. https://doi.org/10.1145/3571730

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. Advances in Neural Information Processing Systems 36. arXiv:2303.13408

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. Journal of Empirical Legal Studies. https://doi.org/10.1111/jels.12413

TextPulse Research. (2026a). Do AI detectors agree? An inter-rater reliability study of commercial AI text detectors on academic writing. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22003419

TextPulse Research. (2026b). Do AI models invent references? A verification audit of citations in AI-generated academic text. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22010511

Topaz, M., Roguin, N., Gupta, P., Zhang, Z., & Peltonen, L.-M. (2026). Fabricated citations: An audit across 2.5 million biomedical papers. The Lancet, 407, 1779-1781. https://doi.org/10.1016/S0140-6736(26)00603-3

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports, 13, 14045. https://doi.org/10.1038/s41598-023-41032-5

Xu, Z., Qiu, Y., Sun, L., Miao, F., Wu, F., Li, X., Wang, X., Lu, H., Zhang, Z., Hu, Y., Li, J., Jin, L., Zhang, F., Luo, R., Liu, X., Li, Y., & Liu, J. (2026). GhostCite: A large-scale analysis of citation validity in the age of large language models. arXiv:2602.06718.