Do AI Models Invent References? A Verification Audit of Citations in AI-Generated Academic Text

TextPulse Research · Working Paper · doi:10.5281/zenodo.22017213

Abstract

Language models are known to invent references, but most published fabrication rates describe models from 2023 and single academic fields, most often being medicine. This study audits reference behavior in five current model families (DeepSeek, Mistral, OpenAI, Anthropic, and Gemini flagship-class models, 2026) under one protocol, across 30 academic topics spanning STEM, medicine, social science, and humanities. It separates the two methods that models use to embed citations in text. In Condition A, we verified 194 author-year citations that these models embedded in text while writing short academic content. After judgement covering books and organization reports, none of the in-text citations was considered a full fabrication. 78.4 percent were fully attributable to a real publication, and the remainder pointed to real authors in the correct field with imperfect year or team detail. In Condition B, each model produced one-shot reference lists with full bibliographic detail, 10 references for each of the 30 topics, yielding 1,500 references in total. Every reference was verified against Crossref and OpenAlex with DOI resolution checks and a judgement pass for thinly indexed works. In this condition, 15.2 percent of references were either fabricated or attached real titles to the wrong authors (95 percent confidence interval 13.5 to 17.1), with family rates from 9.0 percent (OpenAI) to 29.7 percent (Mistral). Another 20.9 percent referred to real works with wrong meta details, such as the DOI or the year. From among the DOIs the models generated, 12.5 percent resolved to a different publication than claimed, while 12.7 percent did not resolve entirely. The contrast between the two conditions is the key finding. The same models that embed almost no fabricated citations in text fabricate at substantial rates when asked for full bibliographic detail, or corresponding reference entries to in-text citations. Every per-reference verdict and its evidence is released as publicly available open data.

1. Introduction

Fabricated references generated by AI chat models have reached academic journal submissions, court filings, and student coursework. Each one transfers verification work to whoever reads the document later, and the volume of that transferred work is now measurable at the scale of the literature itself. An audit of nearly 2.5 million biomedical papers found that one paper in 277 published in early 2026 cited a reference that does not exist, up from one in 2,828 in 2023, a twelve-fold increase whose inflection point in mid-2024 coincides with the high rate of adoption of AI writing tools (Topaz et al., 2026). Fabricated citations have also been confirmed in papers accepted at top AI conferences after full peer review (Russinovich et al., 2026; Ansari, 2026) and in

student coursework (Watson, 2024). Fabricated references are an increasing concern in education and academia.

Models inventing references has been documented since early 2023. In 2026, several questions arise. How often do current flagship AI models still do it? Which kinds of invention remain, whole references that do not exist or real works with corrupted details? Do model families differ enough that the choice of AI tool matters? Has the behavior improved since the studies carried out in 2023?

This study answers these questions with an audit across five model families. It also adds a distinction that the existing literature mostly ignores, the difference between citations embedded in text and references (i.e., bibliographies) produced as lists. The two behaviors face different constraints. An in-text author-year citation carries little detail, and a model can remain safe by naming a real and heavily cited author. A full APA reference commits the model to authors, year, title, venue, volume, pages, and a DOI. Every added field is an additional opportunity to fabricate. Auditing both behaviors for the same models on the same topics shows where along that gradient of detail the invention actually takes place.

2. Background on citations, reference lists, and DOIs

The units this study audits are defined by citation conventions, so we summarize them briefly. Academic attribution operates at two levels. Inside the text, a citation signals that a claim is based on someone else's work. At the end of the document, a reference list entry supplies the full bibliographic record behind each citation. The two levels carry different amounts of information, and this study measures them separately because they fail differently.

In author-year styles such as APA, the in-text citation takes one of two forms. A parenthetical citation places the authors and year in parentheses at the end of a clause, like so "(Slingerland & Chudek, 2011)". A narrative citation makes the authors part of the sentence, with only the year in parentheses, as in "Slingerland and Chudek (2011) found". Both of these forms carry the same minimal payload, surnames and a year. Condition A of this study audits in-text citations of this type.

A reference list entry commits to more details. In APA style in particular, it specifies every author with initials, the year, the full title, the venue, the volume, issue, and page numbers, and a DOI. The conventions also shift over time. The 6th edition of the APA manual listed up to seven authors per entry and printed DOIs as "doi:" strings (American Psychological Association, 2010). The 7th edition lists up to 20 authors, drops publisher locations, presents DOIs as https://doi.org/ links, and shortens in-text citations with three or more authors to "et al." from the first mention (American Psychological Association, 2020). Models trained on a massive volume of academic text have seen both conventions, and the generated lists in this study are a mixture of the two. Condition B audits entries of this type.

The DOI deserves particular attention. A “Digital Object Identifier” is a persistent identifier registered with a central directory, for journal articles almost always through Crossref, and it resolves to the publication it was registered for. Since a DOI is machine-resolvable, it is the strongest integrity signal a reference offers, and the easiest to test automatically. It is also where fabrication becomes apparent. A fabricated reference must either omit the DOI, invent a string that does not resolve, or attach a real DOI that resolves to an entirely different work. Section 5.4 explores all three behaviors, and the third turns out to be common enough to render simple link-checking useless.

3. Literature review

The fabricated reference literature began with audits of ChatGPT in 2023. Walters and Wilder (2023) generated 84 short literature reviews on 42 topics and verified all 636 citations. They found that 55 percent of GPT-3.5 citations and 18 percent of GPT-4 citations were entirely fabricated, and that 43 percent and 24 percent of the real citations respectively carried significant errors. In medicine, Bhattacharyya et al. (2023) verified 115 references from GPT-3.5 medical texts and found 47 percent fabricated and another 46 percent authentic but inaccurate, with only 7 percent fully accurate. Gravel et al. (2023) found that 69 percent of the references ChatGPT generated for 20 medical questions were fabricated, and noted that most fabrications looked real, with real author names and credible venues. Athaluri et al. (2023) found that 28 of 178 generated references could not be traced at all and that many more had broken identifiers.

Subsequent research was on other fields besides medicine and single AI models. Cabezas-Clavijo and Sidorenko-Bautista (2025) assessed 400 references from eight chatbots across five knowledge areas and found only 26.5 percent fully correct, with 39.8 percent fabricated or with errors, and with Grok and DeepSeek outperforming ChatGPT. Linardon et al. (2025) audited 176 GPT-4o citations in mental health literature reviews and found 19.9 percent fabricated and 56.2 percent fabricated or error-ridden, with fabrications more concentrated in less studied topics. The same study reported a finding that anticipates our DOI results, that 64 percent of fabricated citations carrying a DOI carried a real DOI that resolves to an entirely unrelated paper. For benchmarking, the GhostCite evaluation verified 375,440 citations generated by 13 models across 40 domains and found hallucination rates from 14.2 to 94.9 percent, depending on the model (Xu et al., 2026). Reference hallucination has also been documented in commercial systems and deep-research agents, where 5 to 18 percent of cited URLs fail to resolve, and where a simple URL-checking tool reduced non-resolving citations by factors of 6 to 79 (Rao et al., 2026). In the legal domain, Dahl et al. (2024) measured hallucination rates of 58 to 88 percent for general-purpose models on verifiable legal questions. Magesh et al. (2025) showed that even retrieval-based commercial legal research tools hallucinate in 17 to 33 percent of responses. Surveys of hallucination in language models provide the general framework (Ji et al., 2023).

A separate set of works measure how much of this output reaches the published record. Topaz et al. (2026) screened 97.1 million references across nearly 2.5 million biomedical papers and

confirmed 4,046 fabricated citations in 2,810 already published papers, with prevalence increasing twelve times between 2023 and early 2026. Xu et al. (2026) found invalid or fabricated citations in 1.07 percent of 56,381 published AI and security papers, with an 80.9 percent increase in 2025 alone. Russinovich et al. (2026) audited top-tier computer science venues and found that roughly one in twenty NeurIPS and USENIX Security papers contained at least two likely hallucinated references. Also, a taxonomy of 100 fabricated citations that passed NeurIPS 2025 peer review found that two-thirds were total fabrications and about 25 percent were real works with corrupted attributes (Ansari, 2026). Watson (2024) documented the same material arriving in student coursework. Fabricated references are therefore not filtered out by review.

Based on the current literature, three gaps remain unresolved. Published fabrication rates that anchor the public discussion mostly describe outdated 2023 models. Most audits cover one field or one vendor rather than several current families under a single protocol with per-reference open evidence. Also, no audit separates in-text citation behavior from reference-list behavior for the same models and topics, which is the distinction that locates precisely where invention actually occurs. This study aims to address all three gaps.

4. Methods

4.1 Conditions and generation

Both conditions use the 30 topics of our detector-agreement study (TextPulse Research, 2026), which span four discipline groups, 8 STEM, 7 medicine, 8 social sciences, and 7 humanities. They also use the same five model families as that study, namely, deepseek-v4-pro, mistral-large-latest, gpt-5.6-luna, claude-sonnet-5, and gemini-3.1-pro-preview.

Condition A re-uses the 60 texts of that study that contain model-generated writing, 30 fully AI-written passages and 30 hybrid texts whose middle section is AI-written. Every author-year citation that the models embedded while writing was extracted. For hybrid texts, citations already present in the human source passage were excluded by matching against the source, leaving only citations the models explicitly generated. This yields 194 model-generated in-text citations (28 DeepSeek, 53 Mistral, 36 Gemini, 47 Anthropic, 30 OpenAI).

Condition B is a fresh generation by the same above mentioned models. For each topic, each family received the one-shot prompt “List 10 key references for a research paper on [topic]. Provide full bibliographic details in APA style: authors, year, title, journal, volume, pages, and DOI where available. Output only the numbered reference list.” with default settings and no follow-up prompts. Four families were called through their public APIs. This produced 150 lists and exactly 1,500 references.

4.2 Verification

Condition A citations carry no title, so they are scored for attributability against Crossref and OpenAlex. VERIFIED means a real work by the named author(s) in the named year matches the topic area. VERIFIED-AUTHOR-YEAR means a real work by the named team in the named year exists but the topic was not assessed, because methods citations sometimes cross domains. PLAUSIBLE-PARTIAL means a real author publishes in the area but no exact year and team match was found. NO-MATCH means no plausible real work was found. Every NO-MATCH then passed through a judgement step with simple queries, author-only search with one-year tolerance in both indexes, and organization-authored reports such as WHO and IEA publications were tagged as a separate category since scholarly indexes do not cover them reliably.

Condition B references carry full detail, so they are verified strictly. Each reference is parsed into fields. A claimed DOI is resolved through Crossref and classified as valid and matching the claimed work, valid but resolving to a different work, or invalid. The claimed title is first searched in Crossref and, if needed, OpenAlex. REAL-EXACT means the work exists and the fields match. REAL-ERRORS means the work exists but one or more fields are wrong, and the wrong fields are recorded. CHIMERA means a real work's title is attached to the wrong authors or originates from parts of different works. FABRICATED means no matching work was found. FABRICATED and CHIMERA verdicts then passed through the same style of judgement with subtitle-stripped and simple searches, so that thinly indexed books and reports are not miscounted as inventions. Upgrades required a title similarity of at least 0.85.

4.3 Statistics

Rates are reported with Wilson 95 percent confidence intervals. Family differences are tested with a chi-square test on fabricated versus real counts. Discipline-level rates are descriptive. All verdicts, evidence fields, and code are made publicly available.

5. Results

5.1 Condition A results

From among 194 in-prose citations, after judgement, 133 (68.6 percent) were VERIFIED, 19 (9.8 percent) were VERIFIED-AUTHOR-YEAR, 39 (20.1 percent) were PLAUSIBLE-PARTIAL, 3 (1.5 percent) were organization reports outside scholarly indexing, and none remained NO-MATCH. No in-text citation named an author team that did not exist. Families differed primarily in the precision of attribution. Fully attributable shares ranged from 87.2 percent for Anthropic, down to 61.1 percent in the case of Gemini, with the remainder in the plausible-partial category, i.e., real researchers with imperfect year or team details.

The initial strict pass classified 15 citations as NO-MATCH. Every non-organization case resolved to a real work under simple queries. These cases were dominated by well-known books (Zuboff 2019, Whorf 1956, Catford 1965), organization reports (WHO 2021, IEA 2022), and indexing gaps

for well-known articles (Arute et al. 2019). This is a methodological finding in its own right. A purely automatic database-lookup pipeline overcounts fabrication, and the overcount is mostly based on books and reports.

5.2 Condition B results

Of 1,499 parseable references, 958 (63.9 percent) were REAL-EXACT, 313 (20.9 percent) were REAL-ERRORS, 62 (4.1 percent) were CHIMERA, and 166 (11.1 percent) were FABRICATED. The combined fabrication rate, counting FABRICATED and CHIMERA together, was 15.2 percent (95 percent confidence interval 13.5 to 17.1).

Figure F5 shows two references that come from a single Claude list on one humanities topic. One is real in every field. The other names two plausible authors, a plausible title, a real journal, and a DOI that resolves, but the work does not exist and the DOI incorrectly points to the journal's issue-information page. Nothing on the surface separates the two entries. Only resolution against a bibliographic database does.

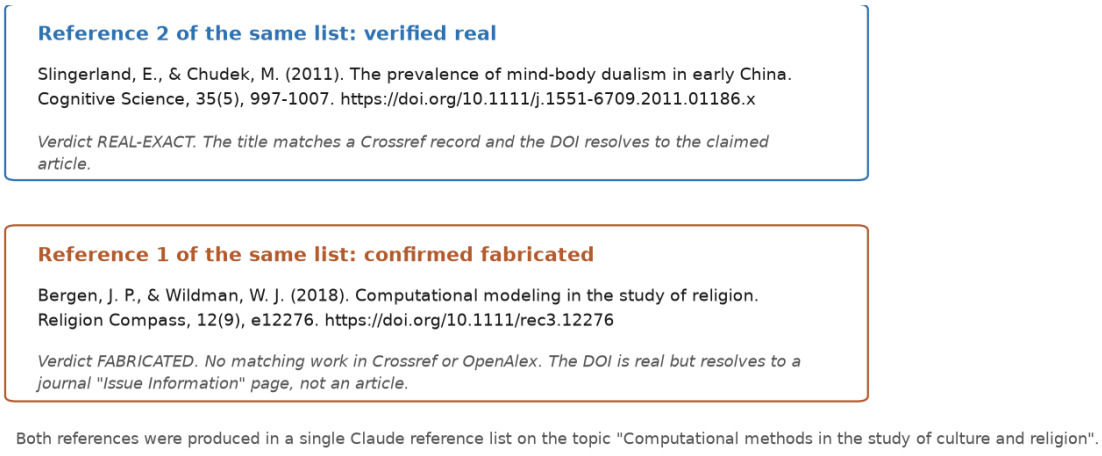

Figure F5. A real and a fabricated reference from the same generated reference list.

Family differences were statistically significant (chi-square 68.1, df 4, n 1,499). OpenAI produced the lowest combined rate at 9.0 percent (confidence interval 6.3 to 12.8), followed by Gemini at 10.3 percent, DeepSeek at 10.7 percent, and Anthropic at 16.3 percent. Mistral produced the highest rate at 29.7 percent (24.8 to 35.1), roughly one fabricated or chimeric reference in every three or four.

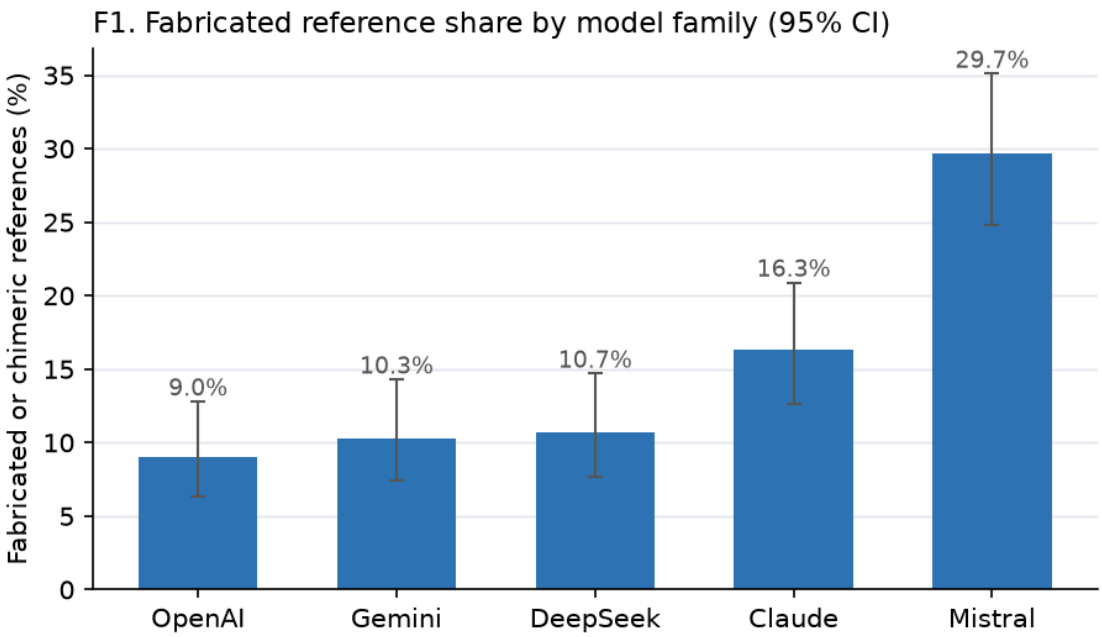

Figure F1. Fabricated reference share by model family with 95 percent confidence intervals.

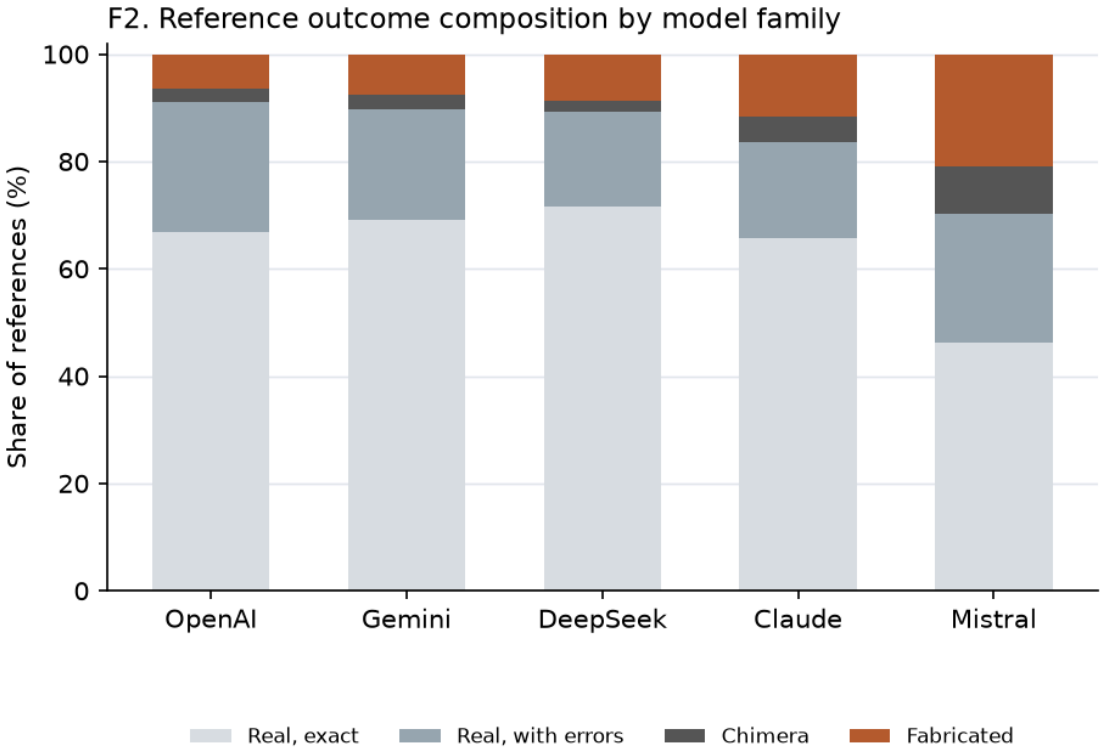

Figure F2. Reference outcome composition by model family.

5.3 Error taxonomy

Among the 313 REAL-ERRORS references, the wrong field was most often the DOI (176 references) or the year (153 references), followed by the journal name (59), the title in a partly matched form (29), and the authors (21). The pattern is consistent across families. Models retrieve the right work and then attach wrong identifying detail to it. Chimera references, most common for Mistral (26) and Anthropic (14), combine a real title with the wrong author team, which makes them resistant to casual checking since both halves of the reference exist.

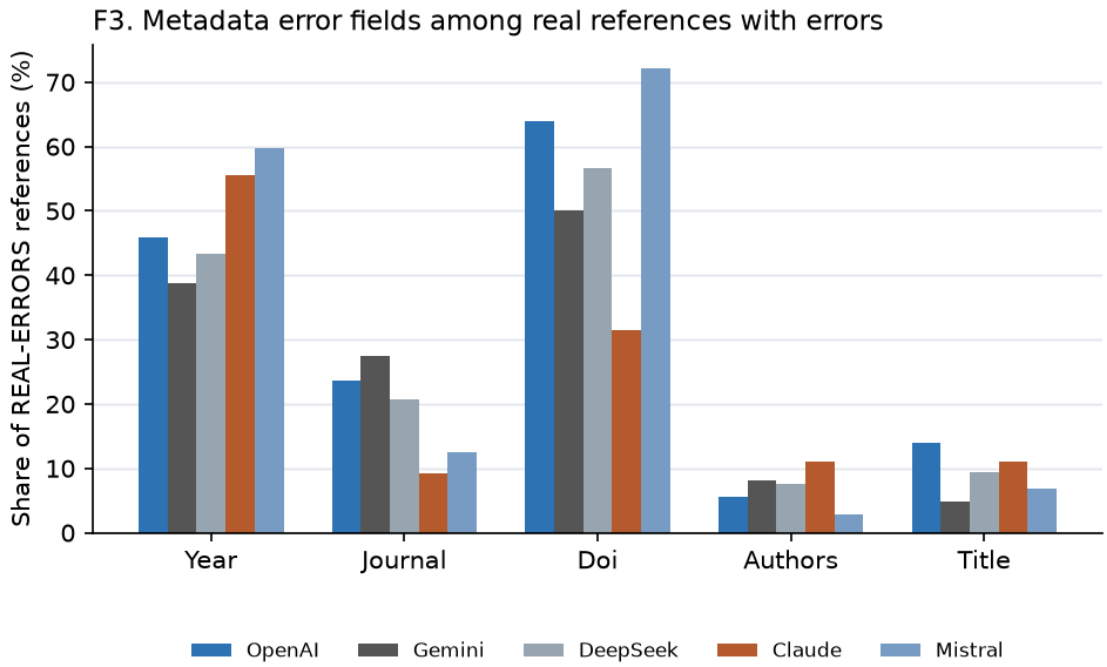

Figure F3. Metadata error fields among real references with errors, by model family.

5.4 DOI validity

The models supplied a DOI for 1,377 of 1,499 references. Of these, 1,030 (74.8 percent) resolved to the claimed work, 172 (12.5 percent) resolved to a different publication than claimed, and 175 (12.7 percent) did not resolve at all. One DOI in four was therefore wrong in a manner that either fails or misleads. The misleading case is a greater concern. A reference whose DOI resolves to a different real paper passes a casual link check while attributing the wrong publication. Mistral again showed the weakest results, with 63 wrong-work DOIs and 65 dead DOIs out of 275 supplied, meaning 46.5 percent of its DOIs were defective. The wrong-work pattern is not specific to this study. Linardon et al. (2025) found that 64 percent of the fabricated GPT-4o citations that had a DOI carried a real one pointing to an entirely unrelated paper. The DOI, the strongest verification signal a reference offers, is the field that models most often fill with convincing but wrong values.

5.5 Disciplines and the contrast between conditions

Fabrication was highest for humanities topics (22.3 percent) and lowest for STEM topics (11.3 percent), with medicine at 14.6 and social science at 13.5 percent. This ordering is consistent with training-data coverage, since the humanities lean on books and less-indexed venues. It also converges with the topic-familiarity effect reported by Linardon et al. (2025) within a single field, where citations on heavily studied topics were almost all real and fabrication concentrated in less-studied ones. The pattern in both of these studies is similar. Models fabricate most where the literature they were trained on is apparently thinnest, which unfortunately is also where human authors are least able to spot an invented reference by inspection.

The contrast between conditions is the clearest single result of this study. The same five models produced 0 outright fabrications in 194 in-prose citations and a 15.2 percent fabrication rate in 1,499 full references, for the same topics. Invention is concentrated where detail is demanded. An author-year mention in prose lets a model stay within well-known names. A full APA reference with a DOI forces commitments the model cannot always hold reliably, and that is the location where fabrication appears.

5.6 Comparison with published 2023 rates

The published 2023 anchors report 55 percent fabrication for GPT-3.5 and 18 percent for GPT-4 on multidisciplinary literature reviews (Walters and Wilder, 2023), 47 percent for GPT-3.5 medical references (Bhattacharyya et al., 2023), and 69 percent for medical questions (Gravel et al., 2023). Intermediate anchors from 2025 fall between the eras. GPT-4o fabricated 19.9 percent of mental health citations (Linardon et al., 2025), and a cross-chatbot audit found 39.8 percent of references erroneous or fabricated across eight systems (Cabezas-Clavijo and Sidorenko-Bautista, 2025). Against these anchors, the 2026 flagship rates of 9.0 to 29.7 percent represent clear but also incomplete improvement, with the warning that protocols differ across studies. The best current families now fabricate at roughly half of the rate of GPT-4 in 2023. The weakest current family in this audit still fabricates at a rate comparable to GPT-4 era models, and the GhostCite benchmark shows that outside the flagship tier the spread remains far wider, from 14.2 to 94.9 percent across 13 models (Xu et al., 2026). Fabrication declined at the top, but it did not disappear. Its form changed to errors that are harder to notice, namely, real works with wrong details and DOIs that resolve to the wrong paper.

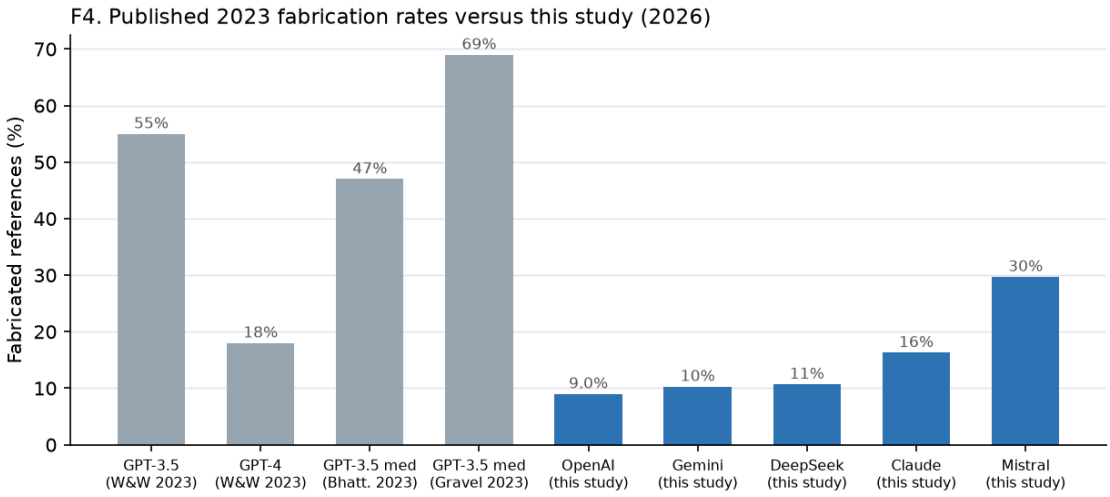

Figure F4. Published 2023 fabrication rates compared with the five families in this study.

6. Discussion

The continuum explains where invention is. In academic content, models cite conservatively and accurately. In reference lists, one reference in about six or seven is fabricated or misattributed, and a further one in five has the wrong details. The practical rule is that any reference list produced by a language model should be treated as unverified input, and every entry should be resolved against a bibliographic database explicitly before actual use. A link that works is not sufficient evidence, since one DOI in eight leads to the entirely wrong publication.

The results also carry a lesson for automated auditing. Our strict pipeline overcounted fabrication in both conditions, in Condition A by misclassifying famous books and organization reports, and in Condition B before an adjudication pass for thinly indexed works. Fabrication studies that relied on a single database lookup without adjudication will overstate rates, and the overstatement falls on the reference types that scholarly indexes cover least well.

The evidence shows the reason the practical rule matters. Peer review does not reliably spot these references. One in twenty papers at NeurIPS and USENIX Security contains at least two likely hallucinated references (Russinovich et al., 2026). A screen of 4,841 accepted NeurIPS 2025 papers confirmed 100 hallucinated citations that three to five expert reviewers had approved (Ansari, 2026). The biomedical literature now absorbs fabricated citations at a measurable and increasing rate (Topaz et al., 2026). Verification therefore must be performed prior to submission. Resolving every DOI and looking up each title in Crossref or OpenAlex is a one-time effort. Also, even a simple resolution check reduces non-resolving citations by an order of magnitude or more (Rao et al., 2026).

For practical developers, the issue can be resolved. The difference between 9 and 30 percent fabrication on identical prompts shows that reference reliability is a property that AI model

vendors can take into consideration, and one that current benchmarks do not present to users.

This study has limitations. Each condition uses one prompt and a single run per family, so the rates describe default one-shot behavior rather than best-case prompting. Topics are only in English. Verification is limited by index coverage even after judgement, and a small number of references may be real works that may not be indexed in Crossref and OpenAlex. The Condition A citation set is modest in size (194 citations) since it is limited by the corpus of the earlier study.

7. Conclusion

Current flagship models no longer invent citations inside text. But when requested for full bibliographic detail, they still invent, at rates from 9.0 to 29.7 percent, based on the family, and they add the wrong details to almost one fifth of their references. The improvement since 2023 is significant. But the fabrication issue has changed in format. References produced by a language model must be checked. AI models save time in finding published works in the form of citations and references, but also cost time in the verification of those works.

Data availability

Per-reference verdicts with evidence for both conditions and all generation, verification, adjudication, and analysis code are openly available on Zenodo (https://doi.org/10.5281/zenodo.22017213). The Condition A corpus texts are already public as part of the detector-agreement study dataset. Full generated texts are not part of the release.

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

Ansari, S. (2026). Compound deception in elite peer review: A failure mode taxonomy of 100 fabricated citations at NeurIPS 2025. arXiv:2602.05930.

Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus, 15(4), e37432. https://doi.org/10.7759/cureus.37432

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.

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. Journal of Legal Analysis, 16(1), 64-93. https://doi.org/10.1093/jla/laae003

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

Linardon, J., Jarman, H. K., McClure, Z., Anderson, C., Liu, C., & Messer, M. (2025). Influence of topic familiarity and prompt specificity on citation fabrication in mental health research using large language models: Experimental study. JMIR Mental Health, 12, e80371. https://doi.org/10.2196/80371

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

Rao, D., Wong, E., & Callison-Burch, C. (2026). Detecting and correcting reference hallucinations in commercial LLMs and deep research agents. arXiv:2604.03173.

Russinovich, M., Siva Kumar, R. S., & Salem, A. (2026). Phantom references: Hallucinated citations that survive peer review at top-tier conferences. arXiv:2607.00738.

TextPulse Research. (2026). 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

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

Watson, A. P. (2024). Hallucinated citation analysis: Delving into student-submitted AI-generated sources at the University of Mississippi. The Serials Librarian, 85(5-6). https://doi.org/10.1080/0361526X.2024.2433640

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.