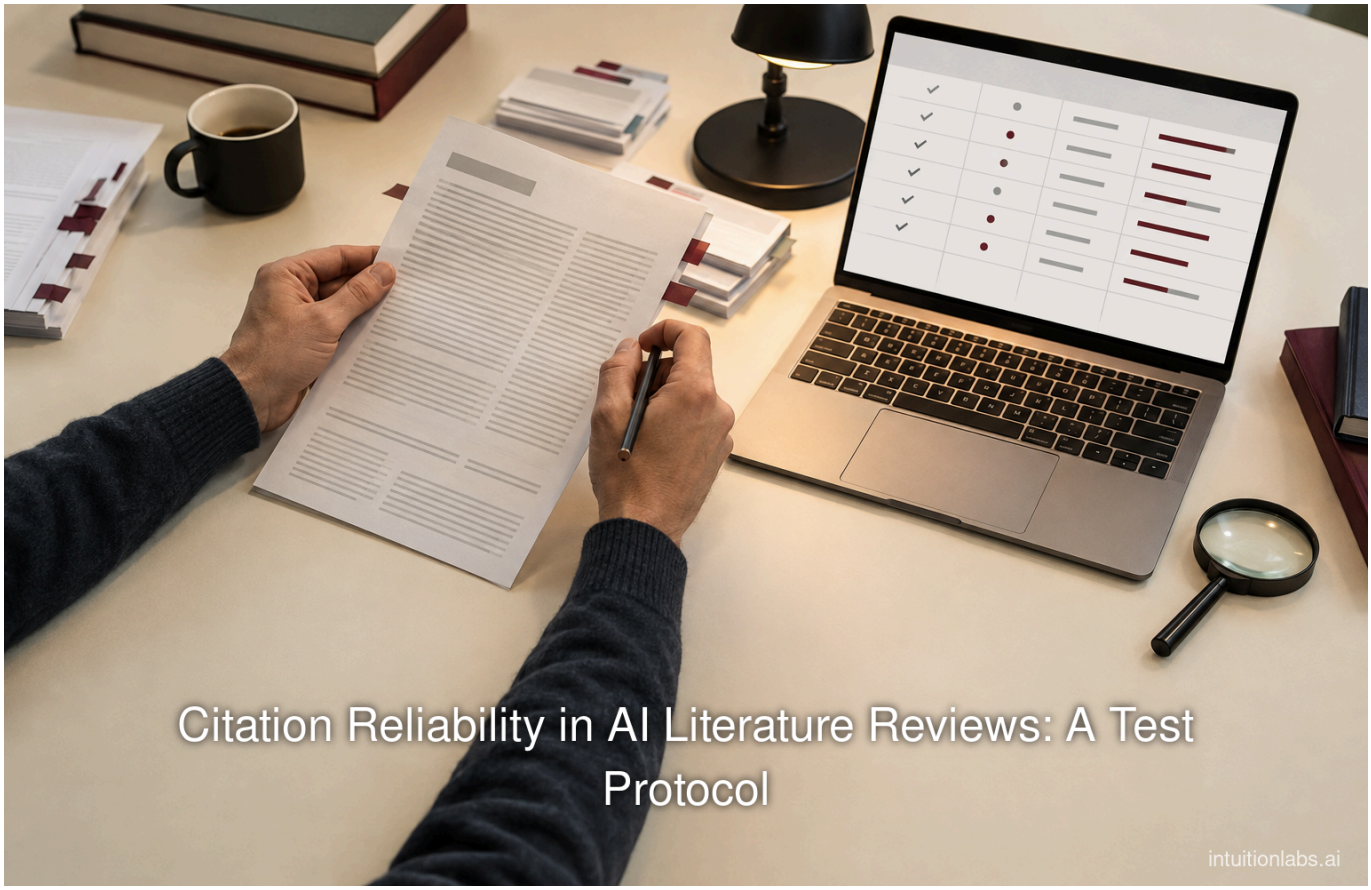
Citation Reliability in AI Literature Reviews: A Test Protocol

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · citation reliability · ai hallucination · literature review · llm citation accuracy · citation verification · ai research tools · hallucinated references · llm benchmarking



Executive Summary

Citation reliability in AI-generated literature reviews is now a measured, published research problem rather than an anecdotal concern, and the measurements available as of September 2026 show wide variation rather than a single answer. Controlled studies find that fabrication rates for citations generated by large language models (LLMs) range from about 5% to over 90% depending on the model and task: a 2023 *Scientific Reports* study found that 55% of citations generated by GPT-3.5 and 18% of citations generated by GPT-4 were fabricated (nature.com), while the unreviewed 2026 GhostCite preprint measured hallucination rates spanning 14.23% to 94.93% across 13 tested models (arxiv.org). Even when a citation is real, it frequently fails a harder test: whether the source actually supports the claim attached to it. A deep-research-agent audit found that leading systems keep link validity above 94% and source relevance above 80%, yet achieve only 39% to 77% factual accuracy on what the cited sources actually say (arxiv.org), showing that retrieval-augmented generation reduces, but does not eliminate, the underlying problem.

This report separates citation reliability into four independently testable dimensions: existence, bibliographic accuracy, claim support, and source relevance, and documents the public verification infrastructure, a free REST API from Crossref, a comprehensive index from OpenAlex, Semantic Scholar's citation-graph API, and the NCBI E-utilities, that a reproducible audit can query without cost (crossref.org) (help.openalex.org). One unreviewed 2026 cross-model preprint reported that requiring agreement across multiple independently queried models before accepting a citation reached 95.6% accuracy, a 5.8-fold improvement over relying on any single model (arxiv.org); this result needs independent replication before it is treated as established practice.

Vendor and independent evidence diverge in a consistent pattern across the commercial tools tested. Elicit, Consensus, and Scite each market citation-grounding guarantees, stating respectively that information is extracted directly from source papers (support.elicit.com), that citations are real and never hallucinated (help.consensus.app), and that every answer is grounded in real, traceable papers (scite.ai). But the one peer-reviewed independent test located in this research, of Scite's citation-classification feature, found the tool's overall classification accuracy was low and less able to distinguish supporting from contrasting citations than human reviewers (journals.indianapolis.iu.edu). General-purpose AI search engines fared worse: an eight-tool independent test by the Columbia Journalism Review's Tow Center found that, collectively, the tools gave incorrect answers to more than 60% of news-citation queries (cjr.org). Meanwhile, researcher adoption has outpaced verification habits: Wiley's 2025 survey of 2,430 researchers found active AI use rose to 84% of researchers in 2025 from 57% a year earlier (wiley.com), even as concern about inaccuracies and hallucinations rose to 64% of researchers from 51% the year before (wiley.com), and a 2024 Ithaka S+R survey of biomedical researchers found that 74% cited insufficient accuracy or reliability as a moderate to large barrier to greater use (sr.ithaka.org).

The evidence supports three conclusions for readers commissioning or relying on AI-assisted literature reviews: fabrication and misattribution rates are model- and domain-dependent enough that no single published number should be treated as universal; automated detection tools currently trade off precision against recall rather than solving citation verification outright, as illustrated by the CiteGuard framework's best result of 68.1% accuracy on a standard citation-attribution benchmark, close to but still short of the 69.2% human baseline (aclanthology.org); and the reproducible four-dimension protocol and free verification infrastructure detailed in this report let any research team test a specific tool or model against its own material rather than relying on vendor assurance alone.

Introduction and Background

Generative artificial intelligence has moved from a novelty for drafting emails to a working tool for literature discovery, and the tool now touches one of the most consequential parts of scholarship: the citation. A citation is the load-bearing element of a literature review. It tells a reader that a specific claim rests on a specific, checkable source. When a **large language model (LLM)** generates that citation, the link between claim and source is no longer guaranteed to be real, and the failure can take several different forms that are easy to conflate.

This report treats **citation reliability** as a measurable property, not an impression, and separates it into four testable dimensions: whether the cited work **exists** at all, whether its **bibliographic details** (author, year, venue, identifier) are correct, whether the source actually **supports the claim** attributed to it, and whether the source is even **relevant** to the surrounding argument. Distinguishing these matters because the same overall "error rate" headline can describe wildly different failure modes: a review that cites 100% real papers but misattributes half of

its claims to them is a different, and arguably more dangerous, problem than one that invents papers outright, because a fabricated title is easy to catch with a database lookup while a misattributed claim on a real, correctly formatted citation often is not.

As of September 2026, the published evidence base for this question has grown substantially. Peer-reviewed audits spanning mental health, computer science, biology, and general-purpose retrieval-augmented generation (RAG) systems have measured citation fabrication rates from single digits to over 90% depending on the model, the task, and the domain, and a parallel literature has emerged proposing standardized error taxonomies and automated verification tools, both detailed in the sections that follow. This report synthesizes that evidence into a single, reproducible protocol; presents the measured error rates by study, model, and tool; and distinguishes vendor claims about citation accuracy from independently verified evidence, so that researchers, librarians, and editors can repeat the audit themselves rather than take any single tool's word for its own reliability.

A Reproducible Protocol for Citation Reliability Testing

Evaluating citation accuracy in a way that other researchers can trust requires a citation audit whose methods are reproducible: it must specify what was measured, on what input set, against which reference database, and with what pass/fail threshold. The literature converges on four distinct test dimensions, which should be scored separately rather than folded into one aggregate "hallucination rate."

The four dimensions, drawn from the taxonomies used across recent audits, are best understood as an escalating chain of checks: a citation must pass existence before bibliographic accuracy is even meaningful to test, and it must pass bibliographic accuracy before claim support can be assessed against the correct source.

- **Existence**: does a work matching the cited title, authors, and venue exist anywhere in an authoritative index. A controlled audit of AI literature-search tools illustrates how much this single check varies by product: in natural-language retrieval mode, one institutional tool returned a correct, resolving digital object identifier (DOI) for only 9 of 40 retrieved records, or 22.5% (mdpi-res.com), while ChatGPT 5.2 in its Extended Thinking mode produced markedly higher DOI-correct rates (mdpi-res.com), reaching 38 of 40 records, or 95.0% (mdpi-res.com), making existence the most basic, and most product-dependent, single check a verification pipeline must run.
- **Bibliographic accuracy**: for a citation that does exist, are the author list, year, venue, and identifier (DOI or arXiv ID) correct. A taxonomy of 100 fabricated NeurIPS 2025 citations separately defines **"Identifier Hijacking"** as a citation that uses a valid scholarly identifier pointing to a real paper, where the authors, title, or other metadata do not actually match the citation text (arxiv.org), a failure mode that is especially dangerous because a reader who only clicks the link will find a real, working page and conclude the citation is sound.
- **Claim support**: does the cited source actually contain evidence for the specific claim attributed to it, independent of whether the citation itself is bibliographically correct. One framework formalizes this as a binary verifier over claim-citation pairs and found the **measured unsupported-citation rate swings from about 3% to about 18%** on identical model output **depending only on the strictness of the automated verifier used to grade it** (arxiv.org), which is itself a finding researchers repeating this protocol need to internalize: the verifier is part of the measurement, not a neutral referee.

- **Source relevance**: even when a citation is real, accurate, and supports a true statement, is it the appropriate source for the specific point being made, as opposed to a tangentially related paper cited because it shares keywords.

Verification infrastructure. A reproducible audit needs machine-checkable reference databases rather than manual lookup. Four public services underpin nearly every technical verification pipeline in the current literature. **Crossref** operates a publicly available REST API that exposes the scholarly metadata deposited by its members (crossref.org), requiring no account since no sign-up is needed and almost none of the metadata is copyrighted (crossref.org); the organization reports driving metadata exchange for more than 25,000 members in 167 countries and supporting 2.1 billion monthly API queries (crossref.org). **OpenAlex** describes itself as a comprehensive index of the world's research (help.openalex.org) and states it indexes more than 320 million scholarly works, with tens of thousands added daily (help.openalex.org). The **Semantic Scholar API** lets a script find and explore data about authors, papers, citations, and venues (semanticscholar.org), over a corpus the service describes as 214 million papers, 2.49 billion citations, and 79 million authors (semanticscholar.org). For biomedical literature specifically, the **NCBI E-utilities** provide a set of nine server-side programs forming a stable interface into the Entrez query and database system (ncbi.nlm.nih.gov), spanning 38 biomedical databases (ncbi.nlm.nih.gov), and include a dedicated **ECitMatch** function built specifically to retrieve PubMed identifiers matching a set of input citation strings (ncbi.nlm.nih.gov), making it the most direct existence-check route for PubMed-indexed claims.

A worked verification pipeline. One unreviewed 2026 methodology preprint operationalized this into a concrete, replicable pipeline that queries Crossref, OpenAlex, and Semantic Scholar in sequence and applies fuzzy title matching to score each citation (arxiv.org), validated against a web-search-enabled LLM as an independent check. A comparable open-source tool, CheckIfExist, uses a similar cascading-database approach and computes a confidence score that weights title similarity against an author-mismatch penalty (arxiv.org), built specifically to catch the kind of author- and title-level mismatches documented throughout this report.

Scoring and reproducibility requirements. To repeat this protocol on a new tool or model: (1) fix a fact sheet of prompts and a target citation count per prompt; (2) run each of the four checks above against Crossref, OpenAlex, Semantic Scholar, or PubMed as appropriate to the field, recording match confidence rather than a binary; (3) report existence, bibliographic accuracy, claim-support, and relevance rates **separately**, with the sample size and the specific verifier used for claim support, since that choice alone can shift the headline number by a factor of six; and (4) publish the raw citation list and verifier outputs so a third party can re-score the sample independently. IntuitionLabs, a life-sciences and AI consultancy, frames the underlying requirement in its own governed-AI implementation work as one of connecting AI systems to sources with **identity, permissions, retrieval, citations, evaluation, logging, and accountable operation** rather than treating citation output as a black box (intuitionlabs.ai).

Measured Citation Error Rates in AI-Generated Literature Reviews

The question "how accurate are AI-generated literature reviews" does not have one answer; it has a distribution, and the distribution is wide. *Table 1* below summarizes measured fabrication and error rates from studies that specifically tested literature-review or citation-generation tasks, as opposed to general-purpose factual question answering.

The earliest widely cited controlled study, published in *Scientific Reports* in 2023, generated 636 references across 84 short literature reviews on 42 topics and found that **55% of GPT-3.5's citations, but just 18% of GPT-4's citations, were fabricated** (nature.com), while even among citations that were not fabricated, **43% of GPT-3.5's real citations and 24% of GPT-4's real citations** contained substantive errors (nature.com). More recent audits show the problem has not disappeared with newer models. A 2026 study of ChatGPT-5-generated mental health literature reviews, auditing 333 citation-claims across 12 reviews, found that **19.8% were judged fully fabricated and a further 18.3% contained major errors**, with only 43.5% fully accurate (link.springer.com) (link.springer.com), and the **proportion judged fabricated ranged from 4% to 37%** across the 12 individual reviews (link.springer.com), showing that a single aggregate number can mask enormous topic-level variance. A separate GPT-4o-specific audit of mental-health reviews found a comparable **19.9% fabrication rate across 176 citations** (ncbi.nlm.nih.gov), and documented markedly **higher fabrication rates for less-visible disorders, such as 28% for binge eating disorder**, than for a highly familiar disorder (ncbi.nlm.nih.gov), suggesting fabrication correlates with how thin the underlying training-data coverage of a topic is.

At the wider benchmark scale, the unreviewed GhostCite preprint tested 13 LLMs and found that, across models, **hallucination rates ranged from 14.23% to 94.93%** (arxiv.org), a study that separately links the trend to the growing use of LLMs in academic writing generally. An unreviewed cross-model preprint built specifically to compare citation reliability across ten commercially deployed systems and four academic domains, covering 69,557 citation instances, found **hallucination rates spanning a fivefold range, from 11.4% to 56.8%** (arxiv.org). Citation-level accuracy also degrades differently from citation-level existence: one large audit of "deep research" agents found these systems **keep link validity above 94% and relevance above 80%, yet achieve only 39% to 77% factual accuracy** (arxiv.org), meaning a citation can point to a real, relevant-looking source while still misrepresenting what that source says. At the level of the reference itself, rather than the claim it supports, a broad URL-based audit found that **3% to 13% of citation URLs are hallucinated**, with no historical record that the page ever existed (arxiv.org). Even peer review is not a full backstop: an automated screen of accepted conference proceedings found that **one in twenty NeurIPS and USENIX Security papers contains at least two likely hallucinated references** (arxiv.org), despite passing expert review.

Table 1. Measured citation and reference error rates by study, task, and model (as of September 2026)

Study	Task and Sample	Reported Rate	Model(s) or Tool(s)
Walters and Wilder, *Scientific Reports*, 2023	636 references across 84 short literature reviews, 42 topics	55% fabricated (GPT-3.5) vs. 18% fabricated (GPT-4)	GPT-3.5 vs. GPT-4
Mental-health literature review audit, 2026	333 citation-claims, 12 LLM-generated reviews	19.8% fully fabricated, 43.5% fully accurate	ChatGPT-5
GPT-4o mental-health citation study	176 citations, 6 disorder-specific reviews	19.9% fabricated overall; up to 28% for less-visible disorders	GPT-4o
GhostCite preprint, 2026	Large-scale citation generation benchmark, 13 models	14.23% to 94.93% hallucination rate	13 commercial and open LLMs
Cross-model reference-audit preprint	69,557 citation instances, 4 academic domains, 10 systems	11.4% to 56.8% hallucination rate	10 commercially deployed LLMs

Study	Task and Sample	Reported Rate	Model(s) or Tool(s)
Deep research agent audit	Multi-agent literature synthesis benchmark	94%+ link validity, only 39-77% factual accuracy	Frontier deep-research agents
URL-based citation audit	Cross-tool citation URL check	3-13% of URLs have no historical web record	Mixed commercial LLMs and RAG tools
Conference proceedings screen	Accepted camera-ready papers, top AI/security venues	1 in 20 papers with 2+ likely hallucinated references	Published human-authored papers (AI-assisted)

The table shows two consistent patterns worth naming directly, both already documented with citations in the paragraphs above: first, fabrication rates for the same underlying task can vary by an order of magnitude depending on the model generation and the vendor's tuning choices, so a single-model spot check cannot stand in for a benchmark; second, the harder-to-detect failure modes (claim support, identifier hijacking) generally show lower headline "rates" than blunt fabrication, but only because most current audits are not yet designed to catch them at scale, not because they are rarer in practice.

Taxonomies of Citation Failure and Detection Tool Performance

Comparing tools and studies requires agreeing on what counts as an error in the first place, and the field has not converged on one taxonomy. Two independent 2026 taxonomies illustrate both the overlap and the disagreement. A study auditing 100 fabricated citations from accepted NeurIPS 2025 papers proposed five categories: **Total Fabrication (66% of cases), Partial Attribute Corruption (27%), Identifier Hijacking (4%), Placeholder Hallucination (2%), and Semantic Hallucination (1%)** (arxiv.org), and found that **all 100 sampled cases exhibited more than one of these failure modes simultaneously**, meaning clean, single-cause errors are the exception rather than the rule. CiteAudit, a separate 2026 benchmark that generated a combined corpus of 2,500 hallucinated and 3,586 verified real citations, instead organizes errors into **Title Errors, Author Errors, and Metadata Errors**, illustrated by its generation method for Title Errors: replacing the original title with a semantically plausible but invalid variant (arxiv.org).

Detection-tool performance varies enormously depending on which of these taxonomies a tool was designed against, and purpose-built multi-agent verification approaches consistently outperform general-purpose LLMs used as ad hoc judges on the same test material. An unreviewed 2026 position-paper preprint from GESIS (Leibniz Institute for the Social Sciences) evaluated five existing open detection tools, including CheckIfExist, HalluCiteChecker, and RefChecker, on a manually verified 104-reference test set and concluded that **their performance is limited by reference extraction errors, incomplete metadata, limited database coverage, and inconsistent verification results** (arxiv.org). The same paper cites large-scale audits estimating that fabricated-reference prevalence in the published literature has risen sharply, from roughly one in 2,828 papers in 2023 to **one in 458 papers in 2025 and one in 277 papers in early 2026** (arxiv.org), a trend line that tracks the mainstreaming of LLM-assisted writing rather than any single tool.

At the citation-attribution end of the taxonomy (finding the correct missing citation for an unattributed claim, rather than checking an existing one), the CiteGuard framework reports its best configuration achieves **up to 68.1% accuracy on the CiteME benchmark**, approaching a 69.2% human baseline (aclanthology.org), on a test set of **130 excerpts collected from human-written manuscripts across different computer science domains**

(aclanthology.org). That headline number obscures a precision-recall trade-off common across this class of tool: CiteGuard's zero-shot baseline configuration achieved near-perfect precision (1.0) but recall as low as **0.17 to 0.29** (aclanthology.org), meaning most detection or attribution tools currently must choose between missing many real errors or flagging too many correct citations as suspect, and no evaluated tool eliminates that trade-off entirely. IntuitionLabs' consultancy work on enterprise AI deployments makes a parallel point in its own hallucination-prevention analysis: retrieval-augmented generation reduces but does not eliminate the underlying problem, which is one reason its prior report on this subject, "AI Hallucinations in Business: Causes and Prevention," which provides an in-depth examination of AI hallucinations in business contexts more broadly (intuitionlabs.ai), frames source-grounding as a mitigation rather than a guarantee.

Benchmarking Large Language Models for Scientific Literature Review

Literature-review-specific audits sit alongside a broader body of general-purpose hallucination and citation-attribution benchmarks that provide comparative context for what "state of the art" currently means. The **Vectara Hallucination Leaderboard**, updated continuously and tracked as of September 2026, measures factual consistency in document summarization, motivated by the observation that **LLMs are increasingly used in RAG and agentic pipelines** where grounded output matters (raw.githubusercontent.com); on that measure, one leading Anthropic model scored a **12.0%** hallucination rate on the benchmark's summarization task (raw.githubusercontent.com), compared with a lower **5.6%** rate for a leading OpenAI model (raw.githubusercontent.com), with the leaderboard's authors cautioning that the method works by **feeding each model the full set of documents and asking it to summarize them** (raw.githubusercontent.com), rather than testing open-domain citation behavior directly.

Citation-specific academic benchmarks predate the current literature-review-specific studies. **ALCE**, the first benchmark proposed specifically for automatic LLM citation evaluation, found that **even the best models lack complete citation support 50% of the time** (arxiv.org), a result that has motivated most of the citation-faithfulness tooling discussed above. **RAGTruth**, a manually annotated corpus of nearly 18,000 RAG-generated responses, demonstrated empirically that **LLMs may still present unsupported or contradictory claims relative to the retrieved content** (arxiv.org), directly undercutting the common assumption that adding retrieval automatically solves the citation-fabrication problem.

Vendor-reported internal benchmarks add a third data point. OpenAI's GPT-5 system card reports that **gpt-5-main's hallucination rate is 26% lower than GPT-4o's, and gpt-5-thinking's is 65% lower than OpenAI o3's** (cdn.openai.com), and separately that **gpt-5-main produces 44% fewer responses with at least one major factual error, and gpt-5-thinking produces 78% fewer**, than the respective prior-generation comparator models (cdn.openai.com), based on an automated factuality grader the company validated by finding **75% agreement with human factuality judgments**, with the grader tending to catch more errors than human raters (cdn.openai.com). Nature's own reporting on the same release noted that OpenAI said **it had reduced the frequency of fake citations and other kinds of hallucination** relative to earlier models (nature.com), without independent replication of the exact figures. At the cross-model level, the 2026 Stanford AI Index Report introduced a new accuracy benchmark and found that **hallucination rates across 26 tested models ranged from 22% to 94%** (hai.stanford.edu), while separately observing that **reporting on responsible-AI benchmarks remains sparse** (hai.stanford.edu) among frontier labs relative to the attention paid to capability benchmarks, a gap this report's own protocol is intended to help close for the specific case of literature-review citations.

Analysis of Key Segments: Tools, Vendor Claims, and Independent Evidence

The commercial AI literature-review and research-synthesis market segments into general-purpose chat/search assistants and purpose-built academic tools, and the gap between vendor claims and independently measured performance differs sharply between the two segments. *Table 2* summarizes representative vendor claims against the independent evidence located for each tool during this audit.

Several purpose-built tools market themselves explicitly on eliminating citation fabrication. Elicit states that it ensures the information it uses is **extracted directly from papers or generated based on research papers, and highlights the source of the content** (support.elicit.com), while separately acknowledging in its own limitations documentation that **the underlying models are not explicitly trained to be faithful to a body of text by default** (support.elicit.com). Consensus goes further, stating that its retrieval architecture means **all citations are real papers, never hallucinated or invented sources** (help.consensus.app), and that of the three hallucination categories it defines internally, **only the third, a misread source, remains possible**, which it says it works to minimize (help.consensus.app), drawing on a database the company describes as covering **more than 220 million peer-reviewed papers** (help.consensus.app). Scite markets a similar guarantee, stating that **every answer is grounded in real papers, never generated or hallucinated** (scite.ai) and that **every claim links back to the specific sentence in the specific paper it came from** (scite.ai). Independent testing of Scite's underlying citation-classification feature, however, found meaningfully different results from the vendor's own tool: one peer-reviewed evaluation reported that **the overall accuracy of scite's assessments was low, and it was less able to classify supporting and contrasting citations than human reviewers** (journals.indianapolis.iu.edu), with the concrete disagreement that **scite classified 2 citations as supporting and 96 as merely mentioning, while the human reviewers found 42 supporting, 39 mentioning, and 17 contrasting** for the same citation set (journals.indianapolis.iu.edu), illustrating that "grounded in real papers" (an existence and relevance claim) is a different guarantee from "correctly classifies what the paper says" (a claim-support guarantee).

Developer-published (not independently replicated) benchmarks show similarly strong self-reported results for PaperQA2, an open-source retrieval agent, whose creators report it **achieves higher accuracy than PhD- and postdoc-level biology researchers at retrieving information from the scientific literature**, as measured on the LitQA2 benchmark (futurehouse.org), and, in the associated peer-reviewed paper, that it **matches or exceeds subject-matter-expert performance on three realistic literature research tasks, without restricting the human comparators** (arxiv.org). Because these figures originate from the tool's own development team rather than a third-party evaluator, they should be read as a strong internal benchmark result pending independent replication, not as an externally audited accuracy figure. A similarly self-published academic search tool, Undermind, states in its own whitepaper that it **misses virtually no highly relevant works found by Google Scholar, while returning ten times as many total relevant results** (undermind.ai), and that it achieves **roughly 98% accuracy in its relevance-classification step** (undermind.ai).

General-purpose AI search engines, by contrast, show consistently weaker independently measured results when tested specifically on citation accuracy for news and public-interest queries. The Columbia Journalism Review's Tow Center tested eight generative search tools, including Perplexity, Gemini, and Grok, and found that collectively

they gave incorrect answers to more than 60% of queries (cjr.org), with the worst-performing tools producing citations where **more than half of the responses from two of the tools cited fabricated or broken URLs leading to error pages** (cjr.org).

Table 2. AI literature-review and research tools: vendor claim versus independent evidence (as of September 2026)

Tool	Vendor Claim	Independent Evidence Found
Elicit	Extracts information directly from papers and highlights the source (support.elicit.com)	Vendor itself notes underlying models are not inherently faithful to source text (support.elicit.com)
Consensus	All citations are real, never hallucinated; 220M+ paper database (help.consensus.app)	No independent third-party accuracy audit located during this research
Scite	Every answer grounded in real papers, traceable to source sentence (scite.ai)	Peer-reviewed test found low agreement with human citation-intent classification (journals.indianapolis.iu.edu)
PaperQA2	Exceeds PhD/postdoc-level accuracy on literature retrieval (LitQA2) (futurehouse.org)	Self-published by developers; not yet independently replicated
Undermind	~98% relevance-classification accuracy; 10x Google Scholar recall (undermind.ai)	Self-published whitepaper; not yet independently replicated
General-purpose AI search (Perplexity, Gemini, Grok, others)	Real-time web-grounded answers with citations	Over 60% of queries answered incorrectly; some tools cited broken or fabricated URLs over half the time (cjr.org) (cjr.org)

The pattern across Table 2 is consistent: every vendor claim located during this research was a claim about **existence and grounding** ("real papers," "never hallucinated"), while the independent evidence, where it existed at all, mostly tested **claim support and classification accuracy**, a different and harder property. Absence of an independent audit for a tool in the table is not evidence that the tool performs well; it reflects a genuine gap in publicly available third-party evaluation for several of these products as of September 2026.

Data Analysis and Evidence

Adoption of AI tools among the researchers who would be affected by citation errors has grown quickly enough that accuracy concerns are now a majority position rather than a minority caution. Wiley's *ExplanAItions* researcher survey, fielded from July 31 to August 18, 2025 with responses from 2,430 researchers around the world (wiley.com), found that active AI use among researchers rose from 57% a year earlier to **84% in 2025** (wiley.com), while concern about inaccuracies and hallucinations rose in parallel, from 51% to **64% of researchers** citing it as a barrier (wiley.com). A separate metric in the same survey found that researchers' confidence in AI's ability to outperform humans has fallen sharply: last year they believed AI already exceeded human ability on 53% of tested use cases; this year they believe that is true for **less than one-third** of tested use cases (wiley.com). A separate 2024 survey of biomedical researchers by Ithaka S+R, drawing on 2,459 total responses including 770 biomedical researchers, found that **63% had experimented with generative AI** (sr.ithaka.org), yet **74% cited insufficient accuracy or reliability** as a moderate to large barrier to greater use, more than any other cited factor

(sr.ithaka.org), and that only **25% used AI to extract knowledge from scientific research** specifically (sr.ithaka.org), showing the accuracy concern is actively suppressing adoption for exactly the use case this report examines.

The scale of hallucinated references entering the published record independently corroborates the survey-based concern. One 2026 cross-repository study spanning arXiv, bioRxiv, SSRN, and PubMed Central, covering 111 million references across 2.5 million papers, produced a **conservative estimate of 146,932 hallucinated citations in 2025 alone** (arxiv.org). A separate audit of proceedings from four high-performance-computing conferences found that **every 2025 proceeding contained citations that could not be verified against any indexed source, affecting 2% to 6% of published papers** (arxiv.org), a marked rise from a near-zero baseline before generative AI writing assistance was in wide use. These figures should be read as **prevalence estimates with stated methodologies, not as a single authoritative count**: each study defines "hallucinated" or "mysterious" slightly differently, and none claims to have audited the entire published literature, so the honest summary is that the direction (a measurable, rising share of citations that fail verification) is well corroborated across independent groups, while the exact percentage is method-dependent and should always be cited alongside its source and sample.

Case Studies and Real-World Examples

(Hypothetical Example) The following worked example illustrates how the four-dimension protocol from this report's methodology section would be applied to a single citation, to make the scoring concrete for a researcher repeating the audit. Suppose an AI-generated literature review includes the sentence: "Prior work has shown that retrieval-augmented generation eliminates citation hallucination in scientific writing (Chen et al., 2024, *Journal of AI Research*)." A reproducible audit would proceed in four steps. First, **existence**: query Crossref, OpenAlex, and Semantic Scholar for a 2024 paper by a first author named Chen in a venue matching "Journal of AI Research"; if no confident match clears the pipeline's threshold, the citation fails at the first gate and the remaining three checks are moot. Second, assuming a plausible match is found, **bibliographic accuracy**: confirm the matched paper's actual authors, year, and venue against the citation string using the weighted title/author/year confidence formula described earlier, since a near-miss match (a real Chen et al. paper in a different, similarly named venue) would fail this step even though the paper exists. Third, **claim support**: retrieve the matched paper's abstract or full text and check whether it actually claims RAG "eliminates" hallucination, as opposed to "reduces" it; the RAGTruth and ALCE findings summarized earlier in this report make an "eliminates" claim inherently suspicious and worth flagging for manual review regardless of what the matched source says. Fourth, **relevance**: confirm the matched paper is actually about RAG and hallucination, not a tangentially related paper about retrieval systems in an unrelated domain that happened to score well on title similarity. A citation that passes existence and bibliographic accuracy but fails claim support, as this hypothetical example is constructed to do, would be scored as a **misattribution error**, a category this report's earlier data sections show is often undercounted by tools that check only for existence.

Implications and Future Directions

The measured evidence in this report points toward three practical implications for anyone commissioning, publishing, or relying on an AI-generated literature review. First, a single hallucination-rate number is not sufficient documentation of reliability; the four-dimension breakdown matters because a tool can score well on existence

while performing poorly on claim support, and vendor marketing materials located in this research consistently emphasized existence-related claims ("grounded in real papers") over claim-support claims, which is precisely the dimension independent testing found weakest. Second, an unreviewed cross-model preprint reports a large accuracy gain when more than three LLMs cite the same work; that promising result should be independently replicated before multi-model consensus is treated as established practice (arxiv.org). Third, the gap this report's protocol section identifies between reporting on capability benchmarks and reporting on responsible-AI or factuality benchmarks, as flagged by the Stanford AI Index's finding that **reporting on responsible-AI benchmarks remains sparse** among frontier labs (hai.stanford.edu), suggests that future model releases are unlikely to close the citation-reliability gap through general capability improvements alone. The improvement seen in OpenAI's own GPT-5 hallucination-rate reduction figures, a **26% to 65% reduction** relative to prior-generation models (cdn.openai.com), indicates progress is real, but the wide, model-dependent spread documented throughout this report's data sections shows the underlying problem remains unsolved industry-wide as of September 2026, not specific to any single vendor. That spread matters most because adoption is rising in parallel: as this report's data section documented, researcher use of AI tools rose to **84% in 2025** even as verification habits lag behind (wiley.com).

For research organizations building AI-assisted literature review into a governed workflow rather than an ad hoc tool, the practical requirement implied by this evidence is closer to the enterprise information-governance framing IntuitionLabs applies to regulated life-sciences AI deployments generally, connecting retrieval, citations, and evaluation into one accountable pipeline rather than trusting a model's citations at the point of generation (intuitionlabs.ai), than it is to a one-time accuracy claim from any single vendor.

Conclusion

Citation reliability in AI-generated literature reviews is measurable, and the measurements available as of September 2026 converge on a consistent, if uncomfortable, picture: fabrication and misattribution rates vary from single digits to over 90% depending on model, task, and domain; no widely used detection tool eliminates the precision-recall trade-off between catching real errors and over-flagging correct citations; and vendor claims of citation accuracy, while often directionally true on the specific dimension of existence, have rarely been independently tested on the harder dimension of claim support. The four-dimension protocol and public verification infrastructure described in this report, built on the same free Crossref and OpenAlex services documented earlier (crossref.org), give researchers, librarians, and editors a reproducible way to test any specific tool or model against their own material rather than relying on either a single vendor's assurance or a single benchmark's headline number. As adoption continues to rise among the research community that would be most affected by these errors, the evidence reviewed here suggests the operative question is no longer whether AI-generated citations should be checked, but which of the four dimensions a given checking method actually covers.

Frequently Asked Questions (FAQs)

How accurate are AI-generated literature reviews?

Accuracy varies enormously by model and task: controlled studies have measured citation fabrication rates from about 5% to over 90% depending on the model and benchmark, with a 2023 controlled study finding 55% of GPT-3.5's citations and 18% of GPT-4's citations fabricated (nature.com); even non-fabricated citations frequently contain claim-support errors, as detailed in the benchmarking section above.

What is the difference between a hallucinated citation and a citation error?

A hallucinated citation typically refers to one that does not exist at all (the existence dimension), while a citation error more broadly includes real citations with wrong bibliographic details, misattributed claims, or irrelevant sourcing, categories this report treats as four separate, separately measurable dimensions.

Can retrieval-augmented generation (RAG) solve citation hallucination?

RAG reduces but does not eliminate the problem. As detailed earlier in this report's data and benchmarking sections, deep-research agents and RAG systems keep link validity high while still misrepresenting what the retrieved sources actually say a substantial share of the time.

How do researchers verify AI-generated citations without specialized software?

University library guidance recommends locating the cited work directly, following any links or the DOI listed in the citation (guides.library.charlotte.edu), since the best way to verify a citation is to find the full text of the source itself (guides.library.charlotte.edu). This is slower than an automated pipeline but requires no tooling beyond a search engine and a database subscription.

Is any single AI literature-review tool citation-error-free?

No tool identified during this research has been independently verified as error-free. Vendor claims of zero hallucination generally address only the existence dimension, and where independent testing exists, such as the peer-reviewed evaluation of Scite's citation classifier, it has found measurable disagreement with human judgment (journals.indianapolis.iu.edu).

What sample size is needed to benchmark a tool's citation accuracy?

Published audits use different sample sizes and designs, so the required sample should be determined for the intended comparison rather than inferred from an aggregate count. The cited 333-citation audit reported descriptive analyses only and did not perform inferential analyses because each study condition had no more than 21 citation-claims (link.springer.com). It therefore does not establish a general minimum sample size or show that 333 aggregate citation-claims are sufficient for topic comparisons.
