Rillor

METHODS NOTE

How to check LLM citations against their sources

A step-by-step method to verify each claim in an LLM answer against the passage it cites, with the citation failure rates that published studies report.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

- Cited LLM answers fail at measurable rates. A 2023 human audit of four generative search engines found 51.5% of sentences fully supported by their citations, and 74.5% of citations supporting their sentence [6]. In a preregistered 2024 study, three commercial legal research tools each hallucinated 17% to 33% of the time [11].
- A citation can fail in five ways, and each needs its own test: the source does not exist or is a different work, one marker covers several claims, the cited passage is the wrong one, the passage does not support the claim, or the model attached the citation without using the passage [1, 13, 19, 22, 28].
- Each step has a published method: identifier lookup with a metadata comparison [5, 9], claim splitting judged on entailment, coverage and decontextualization [15], passage-level entailment with three labels [19] and context ablation for faithfulness [23].
- Automated judges agree with people part of the time and lean lenient. In TREC 2024 an LLM judge matched human support labels 56% of the time and more often gave the higher label [20]. Judges often treat contradicted claims as supported [19]. No source measures the whole check as one pipeline.
- Courts place the check on the person who files. Mata v. Avianca imposed a $5,000 penalty in 2023 [1], and a California appeals court imposed $10,000 in 2025 [3]. A public database counted 2,266 decisions as of 5 November 2026 [29].

Background

Generative systems now answer with in-line citations, and each citation invites the reader to check. Liu, Zhang and Liang set the standard such citations should meet. A system should cite comprehensively, so every statement is fully supported, and accurately, so every citation supports its statement [6]. Their audit of four generative search engines on 1,450 queries in early 2023 found that "a mere 51.5% of generated sentences are fully supported by citations and only 74.5% of citations support their associated sentence" [6]. By engine, the share of sentences fully supported ranged from 11.1% to 68.7% [6]. Citation precision was also inversely correlated with how useful answers seemed (r = -0.96) [6]. An answer that reads well is no evidence that its citations hold.

Other studies measured other failures on other systems.

Study	Domain and sample	Result	Systems from
Liu, Zhang and Liang [6]	1,450 queries, four generative search engines	51.5% of sentences fully supported, 74.5% of citations supportive	Early 2023
Walters and Wilder [9]	636 references in 84 generated literature reviews	55% and 18% fabricated (older and newer model version)	2023
Gao et al., ALCE [7]	Long-form answers on ELI5	About 50% of answers from two widely used model families not fully supported	2023
Dahl et al. [12]	Verifiable questions about federal court cases	Hallucination rates of 58% to 88% across four general-purpose models	2023
Magesh et al. [11]	202 legal research queries	17% to 33% of responses hallucinated, by tool	2024
Wu et al., SourceCheckup [10]	800 medical questions, 58,000 statement and source pairs	50% to 90% of responses not fully supported	2024
Tow Center [31]	1,600 requests to eight generative search tools to identify the news article an excerpt came from (headline, publisher, date and URL)	More than 60% answered incorrectly, 37% to 94% by tool	Early 2025

Study	Domain and sample	Result	Systems from
Venkit et al., DeepTRACE [26]	Deep research configurations	Citation accuracy of 40% to 80%	2025

The errors travel into the record. A May 2026 preprint audited 111 million references across 2.5 million papers and gave "a conservative estimate of 146,932 hallucinated citations in 2025 alone" [27]. Its authors estimate that 78.8% of non-existent citations pass arXiv's moderation, and that 85.3% of those in bioRxiv preprints persist into the published version [27].

NIST's generative AI profile names the risk. It defines confabulation as "The production of confidently stated but erroneous or false content (known colloquially as “hallucinations” or “fabrications”) by which users may be misled or deceived" [4]. Outputs "may also include confabulated logic or citations that purport to justify or explain the system’s answer" [4].

This report is for two groups: teams who receive LLM answers with citations and must decide whether to rely on them, and engineers who build the systems that produce them. Both need the same check, in the same order.

Method

What was examined. The steps that establish that each citation in an LLM answer points to a real source, that the source contains what is quoted, that the cited passage supports the claim and that the model used that passage. For each step: the definition in the research, the methods proposed and how well those methods agree with human judges.

Sources. Thirty-one sources, each opened and read for this report in October 2026: three court decisions from public court copies [1, 2, 3], the NIST profile [4], Free Law Project's citation API documentation [5], journal and conference papers [6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21], preprints [11, 22, 23, 24, 25, 26, 27, 28], a database of court and tribunal decisions and its FAQ [29, 30] and a Tow Center study [31].

How. Each step is described from the papers that define it and set against the validation figures they report. Every rate carries its systems, sample and date. Product and vendor names are left out where a result does not depend on them, and individual lawyers are not named. Nothing here is legal advice.

Findings

1. Confirming the source exists and the quotes match

How sources fail to exist. Walters and Wilder checked the 636 references in 84 literature reviews written by an older and a newer version of a general-purpose chatbot in 2023 [9]. Of the older version's references 55% were fabricated, and 18% of the newer version's [9]. Of the real ones, 43% and 24% had substantive errors such as "incorrect authorship, title, journal, publisher, volume number, pagination, etc." [9]. In medical answers, models without web access produced valid URLs only 40% to 70% of the time [10]. In the Tow Center's 2025 study, one tool's answers to 200 prompts included 154 citations that led to error pages [31].

How to establish existence. Walters and Wilder searched library databases, search engines and publishers' sites. "We regarded a cited work as real (non-fabricated) if we found an actual work that was a match or near-match with regard to both title and author(s)" [9]. Before ruling a work fabricated, they browsed the journal volume or issue and searched the publisher's site [9]. At scale, Zhao and colleagues matched references against an index built from Semantic Scholar and OpenAlex, which matched 95.1% of references [27]. Their counts are an excess of unmatched references over the pre-LLM baseline, "rather than classifications of individual references" [27]. A failed match shows where to look, and a person confirms the individual case.

A working identifier can point to the wrong work. In Mata v. Avianca, the fabricated decision the filing called "Varghese" bore a docket number "associated with the case George Cornea v. U.S. Attorney General, et al." [1]. Its Federal Reporter citation "is associated with J.D. v Azar, 925 F.3d 1291 (D.C. Cir. 2019)" [1]. Ansari's 2026 study of 100 fabricated citations in papers accepted at NeurIPS 2025 calls this identifier hijacking: a valid arXiv ID, DOI or URL points to a real paper whose authors or title do not match [28]. "This creates false verifiability: a reviewer who clicks the link sees a real paper and may assume the citation is correct without verifying the metadata matches" [28]. The 100 citations appeared in 53 accepted papers despite review by three to five experts per paper [28].

Tools resolve citations and leave the comparison to the checker. Free Law Project's citation lookup API "can be useful as a guardrail to help prevent hallucinated citations" [5]. Each result carries a status: found, valid but missing from the archive, a reporter the system does not know, or a match to more than one decision [5]. A text request is limited to 64,000 characters, "about 50 pages of legal content" [5]. A found citation proves that the reporter, volume and page lead to a decision. The checker still compares the case name, court and year there with the citation, because the Mata fabrication carried a real reporter citation [1, 5].

Quotations are a separate test. In Noland v. Land of the Free, "appellant's opening brief contains 23 case quotations, 21 of which are fabrications" [3]. In Al-Haroun, one of the two matters in the joined judgment, the order of Dias J described authorities "many of which appear to be either completely fictitious or which, if they exist at all, do not contain the passages supposedly quoted from them, or do not support the propositions for which they are cited" [2]. That sentence names three failures: a source that does not exist, a quotation it does not contain and a proposition it does not support. Step 1 tests the first two by searching the source text for each quoted string.

The producing tool cannot confirm its own citations. In Mata, counsel asked the chatbot "Is Varghese a real case", and it "responded that it had supplied “real” authorities" [1].

2. Splitting an answer into checkable claims

Sentences hold several claims. Liu and colleagues used sentences "for simplicity and reproducibility", and noted that "a sentence may contain multiple independently-verifiable claims" [6]. A marker after two statements can be read as supporting both or only the last [6]. FActScore breaks text into atomic facts, each "a short sentence conveying one piece of information" [13]. One model family's output held 4.4 pieces of information per sentence, and 40% of those sentences mixed supported and unsupported information [13]. A sentence-level check gives such a sentence one verdict for several claims.

How to split. SAFE uses a model to split a response into facts, revise each to be self-contained, check its relevance and then rate it [14]. Its example turns "It opened in the 20th century" into "The Eiffel Tower opened in the 20th century" [14]. Ragas asks a model to "create one or more statements from each sentence in the given answer" [16]. Rashkin and colleagues evaluate a statement with its context filled in, so "he was 25 years old" is checked as "George Harrison was 25 years old when ‘Wonderwall Music’ was released" [8].

How to judge the split. Metropolitansky and Larson warn that "misrepresenting the source text or omitting factual content can result in misleading or incomplete conclusions" [15]. They set three criteria [15]:

Criterion	Definition
Entailment	"if the source text is true, the extracted claims must also be true"
Coverage	"extracted claims should capture the verifiable information in the source text without explicitly including the unverifiable information"
Decontextualization	Each claim is understandable on its own and keeps its meaning from the source text

Their method extracts claims "only when there is high confidence in the correct interpretation of the source text" [15]. In a citation check, a sentence the splitter cannot interpret goes to a person. Statements about the system itself and questions to the user need no check [6]. A statement with no citation marker is recorded as uncited.

Validation. SAFE agreed with crowdsourced annotators on 72% of about 16,000 facts and was judged right in 76% of 100 sampled disagreements, at more than 20 times lower cost [14]. SAFE rates facts against web search [14]. A citation check keeps its splitting step and rates each claim against the cited passage.

3. Matching each claim to a retrieved passage

The attribution test. Rashkin and colleagues define a statement as attributable when "A generic hearer will, with a chosen level of confidence, affirm the following statement: “According to P, s”" [8]. Liu and colleagues apply it to citations: a statement is fully supported if a generic hearer would affirm "According to cited webpages Ci, si" [6].

Attribution is a narrower question than truth. The framework avoids "absolute judgments regarding “factuality” of utterances" and leaves source quality to a separate evaluation [8]. A true claim absent from its cited passage fails the check. A claim drawn faithfully from a wrong source passes it, and needs a review of source quality [8]. The framework first asks whether a statement is understandable without the source, and only then whether "all of the information provided by the system response" is "fully supported by the source document" [8].

Misgrounding. Magesh and colleagues define the failure: "A response is misgrounded if key factual propositions are cited but misinterpret the source or reference an inapplicable source" [11]. They count a response as hallucinated "if it is either incorrect or misgrounded" [11]. One tool gave a summary that was "the direct opposite of the actual holding of a case", and "Identifying these misunderstandings often requires close analysis of cited sources" [11].

Passage length. The TRUE study found that NLI and question-answering approaches to factual consistency do well across 11 datasets, and that "all approaches struggle with long inputs, labeling issues and personal statements" [17]. TREC 2024 assessors read each answer sentence against a passage of typically 500 to 1,000 characters, and participants could cite up to 20 passages per sentence [20]. Matching a claim to a whole document gives the checker more text than these studies used.

The passage should be one the system retrieved. Wallat and colleagues define a faithful citation as one whose document is in the set the model retrieved, supports the statement and causally affected it [22]. The first condition is checkable from logs. A system that records every retrieved passage can flag any citation to a document it never retrieved.

4. Entailment checks with NLI models or LLM judges

Labels. Yue and colleagues sort each pair of claim and reference into three types: attributable when the reference fully supports the statement, extrapolatory when it lacks sufficient information, and contradictory when the statement directly contradicts it [19]. Liu and colleagues ask for "full, partial, or no support" [6], and TREC 2024 uses the same three levels [20]. No source sets a standard for partial support or a score threshold for counting a claim as supported.

NLI models. ALCE uses the TRUE NLI model to test whether cited passages entail each statement [7]. Against human labels it showed "substantial agreement for citation recall (0.698) and moderate agreement for citation precision (0.525)", with accuracy of 85.1% and 77.6% [7]. Its authors note that "the NLI model cannot

detect the case of “partially support””, which lowers measured precision [7]. Tang, Laban and Durrett trained a 770-million-parameter checker that reaches a large frontier model's accuracy on their LLM-AggreFact benchmark at 400 times lower cost [18].

LLM judges. The TREC 2024 RAG track compared an LLM judge with human assessors across 45 submissions on 36 topics [20]. Labels matched exactly for 56% of assessments made from scratch and 72% when humans post-edited the judge's labels [20]. Run-level scores correlated above 0.79 Kendall tau [20]. The judge was "more likely to provide a higher support label than the human judge", and in 15.1% of cases it said partial support where the human said none [20]. An independent human judge agreed more with the LLM judge than with the original assessor, which the authors read as support for LLM judges [20]. Run rankings held up, and single labels varied.

Where judges fail. Yue and colleagues name "insensitivity to fine-grained information comparisons, such as overlooking contextual cues in the reference, disregard for numerical values, and failure in performing symbolic operations" [19]. Prompted and fine-tuned models "often fail when faced with contradictory inputs, most often treating them as attributable" [19]. A contradicted claim marked as supported passes the check while saying the opposite of its source.

Claim-level scoring. Ragas scores faithfulness as the share of extracted statements an LLM finds supported by the retrieved context, and asks for "a brief explanation for each statement before arriving at the verdict (Yes/No)" [16].

Check	Validation reported	Source
ALCE NLI citation recall and precision	Kappa 0.698 and 0.525, accuracy 85.1% and 77.6%	[7]
TREC 2024 LLM judge, three levels	56% exact match from scratch, 72% with post-editing	[20]
Ragas faithfulness	95% agreement on 50 Wikipedia pages	[16]
SourceCheckup	89% agreement with three medical experts' consensus	[10]
SAFE, facts against web search	72% agreement on about 16,000 facts	[14]

Each figure comes from a different task and dataset, so the rows do not rank the methods.

Who judges. Panickssery, Bowman and Feng found self-preference, "where an LLM evaluator scores its own outputs higher than others' while human annotators consider them of equal quality" [24]. The study measured general quality ratings. It still supports using a judge that differs from the model that wrote the answer.

5. Faithfulness: citations attached after the answer was written

Correctness and faithfulness differ. Wallat and colleagues separate correctness, "whether cited documents support the corresponding statements", from faithfulness, "whether the citation accurately depicts the model's reasoning process" [22]. The gap is post-rationalization, "where models may cite sources to fit preconceived notions rather than genuine retrieval" [22].

The experiment. They gave a 104-billion-parameter retrieval-tuned model NaturalQuestions questions with five retrieved documents [22]. They appended statements from its cited answers to other documents and asked again. Where the original statement reappeared, the model cited the altered document in 12% of cases for random documents (116 of 936), 57% for relevant documents it had not cited (273 of 476) and 55% for documents cited for another reason (290 of 525) [22]. The authors "conjecture post-rationalization to be a common phenomenon" and say their "empirical analysis is limited" [22]. The figures come from one model and one dataset in a December 2024 preprint. A later version in the ICTIR 2025 proceedings was not read for this report.

Post-hoc citing by design. RARR "automatically finds attribution for the output of any text generation model" and post-edits it [21]. Such a design can produce correct citations for text that came from elsewhere. On ALCE, a closed-book model "with post-hoc citing achieves good correctness but much worse citation quality", because its text is often "not similar to any retrieved passages" [7].

Testing faithfulness. Cohen-Wang and colleagues distinguish two kinds of attribution. "Contributive attribution identifies the sources that cause a model to generate a statement. Meanwhile, corroborative attribution identifies sources that support or imply a statement" [23]. Steps 3 and 4 are corroborative. ContextCite is contributive: it ablates sources and tracks the probability of the original response, so it needs the model's output probabilities [23]. When a model misreads its context, "A corroborative method might not find any attributions (because nothing in the context supports its statement)", while a contributive method finds the misread passage [23]. Without probabilities, a test from Wallat and colleagues runs from outside: "Adding irrelevant documents to the context should not affect the attribution, provided that the answer remains unchanged" [22].

6. What happens to a claim that fails

Revise against the evidence. RARR "first runs an agreement model to detect disagreement between the text and the evidence, then runs an edit model to revise the text if needed" [21]. Its example corrects a marathon world record from 2:01:39 in 2018 to 2:01:09 in 2022 [21]. A revised claim goes through steps 3 and 4 again.

Check independently. Chain-of-Verification answers verification questions "independently so the answers are not biased by other responses" [25]. The checker should not see the draft's reasoning when it judges a claim.

Another model pass is no substitute for reading. In Mata the chatbot vouched for its own fabrications [1]. In Noland, counsel ran the briefs "through other AI platforms to check for errors" and admitted he "did not read the “enhanced” briefs before he filed them" [3]. Magesh and colleagues conclude that "Users of these tools must continue to verify that key propositions are accurately supported by citations" [11].

Remove or escalate. A claim no retrieved passage supports, and that revision cannot ground, comes out of the answer or goes to a person with its passage and label. NIST lists the action to "Review and verify sources and citations in GAI system outputs during pre-deployment risk measurement and ongoing monitoring activities" [4]. It also warns: "Avoid extrapolating GAI system performance or capabilities from narrow, nonsystematic, and anecdotal assessments" [4]. A record of each claim, citation, passage, label and action gives monitoring its data.

7. Measuring citation precision and recall

Definitions. Liu and colleagues define citation recall as "the proportion of verification-worthy statements that are fully supported by their associated citations", and precision as "the proportion of generated citations that support their associated statements" [6]. They combine the two as citation F1, the harmonic mean [6]. Precision guards against over-citing: "a response that cites every webpage on the Internet for each generated statement would have high citation recall, but low citation precision" [6]. ALCE counts a citation as irrelevant if it cannot support the statement and removing it leaves the other citations' support intact [7]. It does not require a minimal set, "because human writing often cites redundant sources to enhance credibility" [7]. TREC 2024 weights both measures, "where precision penalizes overcitation, and recall penalizes undercitation" [20].

Statement level and response level. SourceCheckup separates three measures [10]:

Metric	Definition
Source URL validity	"the proportion of generated URLs that return a valid webpage"
Statement-level support	"the percentage of statements that are supported by at least one source in the same response"
Response-level support	"the percentage of responses that have all their statements supported"

For one retrieval-enabled model on consumer health questions, URL validity was 100%, statement-level support 75.7% and response-level support 38.4% [10]. DeepTRACE found citation accuracy of 40% to 80% across deep research systems in 2025, and notes that "Listing more sources does not guarantee better grounding" [26].

8. What courts have done with unchecked citations

Mata v. Avianca (S.D.N.Y., 2023). The court searched for the cited cases itself "but was unable to locate multiple authorities", and the Clerk of the Eleventh Circuit confirmed that one decision "is not an authentic ruling of the Court" [1]. The opinion's first paragraph says: "Technological advances are commonplace and there is nothing inherently improper about using a reliable artificial intelligence tool for assistance. But existing rules impose a gatekeeping role on attorneys to ensure the accuracy of their filings" [1]. The respondents "continued to stand by the fake opinions after judicial orders called their existence into question" [1]. The court imposed a penalty of $5,000, "jointly and severally imposed on Respondents" [1].

Ayinde and Al-Haroun (High Court, Divisional Court, 2025). Freely available generative AI tools "are not capable of conducting reliable legal research", and those who use them "have a professional duty therefore to check the accuracy of such research by reference to authoritative sources" [2]. In Al-Haroun, the schedule "lists forty five citations that had been put before the court. In eighteen instances, the case cited does not exist" [2]. "A lawyer is not entitled to rely on their lay client for the accuracy of citations of authority or quotations" [2].

Noland v. Land of the Free (California Court of Appeal, 2025). The court published its opinion "as a warning" [3]. No paper filed in any court "should contain any citations", from generative AI or any other source, "that the attorney responsible for submitting the pleading has not personally read and verified" [3]. It imposed "a conservative sanction of $10,000" and directed service of the opinion on the State Bar [3]. It declined to order sanctions payable to the other side, because "respondents did not alert the court to the fabricated citations" [3].

The database. Damien Charlotin's database counted 2,266 decisions as of 5 November 2026 [29]. It covers cases where a court "has explicitly found (or implied) that a party relied on hallucinated content or material", with some exceptions where use of a generative model was alleged and not confirmed [29].

Filter on the database page, 5 November 2026	Count
Fabricated	1,856
False quotes	602
Misrepresented	960
Outdated advice	38
Party: pro se litigant	1,305
Party: lawyer	898

Source: [29].

A decision can fall in more than one category, so the four categories sum to 3,456 against 2,266 decisions. The author estimates two or three cases a month before 2025 and about five a day by February 2026, and calls the database "necessarily an undercount" [30]. Dahl and colleagues found hallucination rates of 58% to 88% across four general-purpose models on verifiable questions about federal cases in 2023 [12].

Implications

For teams receiving LLM answers. Test each citation in order. Confirm the source exists and is the one named, by identifier and metadata [1, 5, 28]. Find each quotation in the source text [2, 3]. Split the answer into claims and attach each to its passage [13, 15]. Judge support with three labels, and send contradictions and partial support to a person [19, 20]. Ask how citations were produced, since citations added after the text can be correct and unfaithful [7, 22]. Courts in the United States and England expect the person who relies on a citation to read it [1, 2, 3].

For engineers building retrieval systems with citations. Log every retrieved passage and require each citation to point to one [22]. Split answers into self-contained claims [14, 15] and match them to short passages [17, 20]. Use a judge that differs from the generator [24], and validate it on labeled cases with contradictions and numbers [19]. Report recall and precision at the statement and response level [6, 10]. Test faithfulness by ablation where probabilities are available [23], and with planted documents where they are not [22]. NIST expects this review before deployment and in monitoring [4]. Rillor builds agentic systems on a client's documents, records and data. In each system, an audit step compares every citation in an agent's findings with the logged evidence, and a claim without matching evidence goes back for rework or is removed.

For teams evaluating LLM research tools. Vendor claims about eliminating hallucinations in legal research were tested and found overstated [11]. Ask for citation recall and precision on your own documents, any automated judge's agreement with human reviewers, and the share of answers fully supported, which can sit far below the share of statements supported [10].

Limits of this analysis

- No source measures the full check, from existence to faithfulness, as one pipeline. Each step is validated on its own data, and the error rates cannot be combined into one catch rate.
- The failure rates describe systems from 2023 to 2025. No answer-level audit here postdates DeepTRACE, from September 2025 [26]. The May 2026 study measures references in papers [27]. No current rate for any model family is given.
- Agreement figures are task-specific [10, 14, 16, 20]. None reports a false-pass rate on business or legal documents.
- Evidence on faithfulness is thin: one model on one dataset [22], and ablation needs model probabilities [23].
- Wallat and colleagues [22] and Magesh and colleagues [11] are cited from preprints. Their published versions (ICTIR 2025 and the Journal of Empirical Legal Studies, 2025) could not be opened.
- The self-preference study measured general quality ratings, not citation checks [24].
- No source gives the time or cost of checking a citation by hand.
- The court database is an undercount by its author's account, and its trend figures are his estimates [29, 30].

Sources

1. Mata v. Avianca, Inc., No. 22-cv-1461 (PKC), Opinion and Order on Sanctions, ECF No. 54. U.S. District Court for the Southern District of New York, 22 June 2023. https://storage.courtlistener.com/recap/gov.uscourts.nysd.575368/gov.uscourts.nysd.575368.54.0.pdf

2. R (Ayinde) v London Borough of Haringey, and Al-Haroun v Qatar National Bank QPSC, [2025] EWHC 1383 (Admin). High Court of Justice, King's Bench Division, Divisional Court, 6 June 2025. https://www.judiciary.uk/wp-content/uploads/2025/06/Ayinde-v-London-Borough-of-Haringey-and-Al-Haroun-v-Qatar-National-Bank.pdf

3. Noland v. Land of the Free, L.P., B331918, 114 Cal.App.5th 426. Court of Appeal of the State of California, Second Appellate District, 12 September 2025. https://storage.courtlistener.com/pdf/2025/09/12/noland_v._land_of_the_free_l.p..pdf

4. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology, July 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

5. Citation Lookup and Verification API. Free Law Project, CourtListener, viewed 24 October 2026. https://www.courtlistener.com/help/api/rest/citation-lookup/

6. Evaluating Verifiability in Generative Search Engines. Findings of EMNLP 2023, arXiv 2304.09848 v2, 23 October 2023. https://arxiv.org/abs/2304.09848

7. Enabling Large Language Models to Generate Text with Citations. EMNLP 2023, arXiv 2305.14627 v2, 31 October 2023. https://arxiv.org/abs/2305.14627

8. Measuring Attribution in Natural Language Generation Models. Computational Linguistics 49(4), December 2023, arXiv 2112.12870 v2, 2 August 2022. https://arxiv.org/abs/2112.12870

9. Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports 13, 14045, 7 September 2023. https://www.nature.com/articles/s41598-023-41032-5

10. An automated framework for assessing how well LLMs cite relevant medical references. Nature Communications 16, 3615, 16 April 2025. https://www.nature.com/articles/s41467-025-58551-6

11. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. arXiv preprint 2405.20362 v1, 30 May 2024, later published in the Journal of Empirical Legal Studies 22(2), 2025. https://arxiv.org/abs/2405.20362

12. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis 16(1), 2024, arXiv 2401.01301 v2, 21 June 2024. https://arxiv.org/abs/2401.01301

13. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. EMNLP 2023, arXiv 2305.14251 v2, 11 October 2023. https://arxiv.org/abs/2305.14251

14. Long-form factuality in large language models. NeurIPS 2024, arXiv 2403.18802 v4, 7 November 2024. https://arxiv.org/abs/2403.18802

15. Towards Effective Extraction and Evaluation of Factual Claims. ACL 2025, arXiv 2502.10855 v2, 6 June 2025. https://arxiv.org/abs/2502.10855

16. Ragas: Automated Evaluation of Retrieval Augmented Generation. EACL 2024 System Demonstrations, arXiv 2309.15217 v2, 28 April 2025. https://arxiv.org/abs/2309.15217

17. TRUE: Re-evaluating Factual Consistency Evaluation. NAACL 2022, arXiv 2204.04991 v3, 3 May 2022. https://arxiv.org/abs/2204.04991

18. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. EMNLP 2024, arXiv 2404.10774 v2, 1 October 2024. https://arxiv.org/abs/2404.10774

19. Automatic Evaluation of Attribution by Large Language Models. Findings of EMNLP 2023, arXiv 2305.06311 v2, 7 October 2023. https://arxiv.org/abs/2305.06311

20. Support Evaluation for the TREC 2024 RAG Track: Comparing Human versus LLM Judges. SIGIR 2025, arXiv 2504.15205 v1, 21 April 2025. https://arxiv.org/abs/2504.15205

21. RARR: Researching and Revising What Language Models Say, Using Language Models. ACL 2023, arXiv 2210.08726 v3, 31 May 2023. https://arxiv.org/abs/2210.08726

22. Correctness is not Faithfulness in RAG Attributions. arXiv preprint 2412.18004 v1, 23 December 2024, later published in the ICTIR 2025 proceedings. https://arxiv.org/abs/2412.18004

23. ContextCite: Attributing Model Generation to Context. arXiv preprint 2409.00729 v2, 13 September 2024. https://arxiv.org/abs/2409.00729

24. LLM Evaluators Recognize and Favor Their Own Generations. arXiv preprint 2404.13076 v1, 15 April 2024. https://arxiv.org/abs/2404.13076

25. Chain-of-Verification Reduces Hallucination in Large Language Models. arXiv preprint 2309.11495 v2, 25 September 2023. https://arxiv.org/abs/2309.11495

26. DeepTRACE: Auditing Deep Research AI Systems for Tracking Reliability Across Citations and Evidence. arXiv preprint 2509.04499 v1, 2 September 2025. https://arxiv.org/abs/2509.04499

27. LLM hallucinations in the wild: Large-scale evidence from non-existent citations. arXiv preprint 2605.07723 v1, 8 May 2026. https://arxiv.org/abs/2605.07723

28. Compound Deception in Elite Peer Review: A Failure Mode Taxonomy of 100 Fabricated Citations at NeurIPS 2025. arXiv preprint 2602.05930 v1, 5 February 2026. https://arxiv.org/abs/2602.05930

29. AI Hallucination Cases Database. Damien Charlotin, last updated 5 November 2026, viewed 7 November 2026. https://www.damiencharlotin.com/hallucinations/

30. Hallucinations Case Database: FAQ. Damien Charlotin, Artificial Authority, 14 November 2025. https://artificialauthority.ai/p/hallucinations-case-database-faq

31. AI Search Has a Citation Problem. Columbia Journalism Review, Tow Center for Digital Journalism, 6 March 2025. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php

About this report

Published by Rillor on 7 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. How to check LLM citations against their sources. 7 November 2026. https://rillor.com/insights/checking-llm-citations-against-sources

Notices

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Tell us the workflow, the sources it draws on and the decisions it supports, and we scope the agents, tools and tests around them. Contact us at rillor.com/contact?type=agentic-systems or contact@rillor.com. More on agentic systems: rillor.com/si/agentic-systems