Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Hallucinated Legal Citations: Lessons from Mata v. Avianca

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-29 · mata v avianca · ai hallucinated case citations · legal ai hallucination rate · lawyers chatgpt fake cases · legal research ai reliability

Original article: https://reliablecontext.com/articles/hallucinated-legal-citations-mata-v-avianca-controls-that-work



In this article

1. Why this case still matters
2. What happened in Mata v. Avianca
3. What the research studies measured
4. Why plausible citations are the dangerous failure
5. Control 1: resolve every citation to a retained primary source
6. Control 2: passage-level support check
7. Control 3: a declared "not found"
8. A worked example
9. Court rules and disclosure
10. What the research says about evaluation
11. Checklist
12. Summary
13. Sources

Why this case still matters

On 22 June 2023, Judge P. Kevin Castel of the US District Court for the Southern District of New York sanctioned two lawyers and their firm for filing a brief that cited judicial opinions that did not exist. The opinions had been produced by ChatGPT. *Mata v. Avianca, Inc.*, No. 22-cv-1461 (PKC), has since become the standard reference point whenever a court, a bar or a firm discusses AI and legal research.

The case is often summarised as "a lawyer trusted ChatGPT". The court record supports a more useful reading. The failure was not that a tool produced false text; it was that a filing process had no step that resolved each citation to a primary source, and that when the opposing party and the court raised doubts, the response did not either. This article reconstructs what happened from the court's opinion, summarises what two empirical studies measured, and derives three controls. It is educational and is not legal advice.

What happened in Mata v. Avianca

The facts below are taken from the court's *Opinion and Order on Sanctions*, Document 54, filed 22 June 2023 (Mata v. Avianca, Inc., ECF 54; docket at CourtListener).

Date (2023)	Event, as found by the court
1 March	Plaintiff's counsel files an affirmation in opposition to a motion to dismiss that cites the fake cases
15 March	Avianca's reply states that counsel "has been unable to locate most of the case law cited" and that "the few cases which the undersigned has been able to locate do not stand for the propositions for which they are cited"
11 and 12 April	The court orders production of the cited cases
25 April	Counsel files an affidavit annexing purported copies of the opinions

Date (2023)	Event, as found by the court
25 May	After an order to show cause, counsel begins, in the court's words, to "dribble out the truth"
8 June	Sanctions hearing
22 June	Opinion and Order on Sanctions

The opening paragraph of the opinion frames the issue precisely: "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." The respondents "abandoned their responsibilities when they submitted non-existent judicial opinions with fake quotes and citations created by the artificial intelligence tool ChatGPT, then continued to stand by the fake opinions after judicial orders called their existence into question."

Three findings are directly relevant to controls.

1. **The research tool was a substitute, not a supplement.** The firm used Fastcase, whose account "had limited access to federal cases". The lawyer testified that he "falsely assumed" ChatGPT "was like a super search engine" (Findings of Fact; transcript quoted in ECF 54).
2. **Verification was asked of the same tool.** The lawyer's affidavit attached screenshots in which he asked ChatGPT "Is Varghese a real case" and "Are the other cases you provided fake"; ChatGPT "responded that it had supplied 'real' authorities that could be found through Westlaw, LexisNexis and the Federal Reporter" (Finding 45).
3. **The fabricated opinions were plausible on the surface.** The court found that the fake "Varghese" decision "shows stylistic and reasoning flaws that do not generally appear in decisions issued by United States Courts of Appeals. Its legal analysis is gibberish" (Finding 27). Yet it carried a docket-style citation and named real judges.

The court found "bad faith on the part of the individual Respondents based upon acts of conscious avoidance and false and misleading statements to the Court", imposed sanctions under Rule 11 of the Federal Rules of Civil Procedure "or, alternatively, its inherent authority", and held the firm jointly responsible. The sanctions were a penalty of $5,000, "jointly and severally imposed", and letters to the plaintiff and to "each judge falsely identified as the author of the fake 'Varghese', 'Shaboon', 'Petersen', 'Martinez', 'Durden' and 'Miller' opinions" (docket entry 54).

The opinion also lists the harms: the opposing party "wastes time and money in exposing the deception", the court's time is diverted, the client "may be deprived of arguments based on authentic judicial precedents", and the reputations of judges "whose names are falsely invoked" are harmed.

What the research studies measured

Two studies from the same research group are the most cited empirical evidence on legal hallucination. They measure different things and should not be quoted interchangeably.

	Dahl et al., *Large Legal Fictions*	**Magesh et al., *Hallucination-Free?***
Systems tested	General-purpose LLMs (including ChatGPT 4 and Llama 2)	Legal research tools: Lexis+ AI, Westlaw AI-Assisted Research, Ask Practical Law AI; GPT-4 for comparison
Task	"Specific, verifiable questions about random federal court cases"	A preregistered set of 202 legal research queries
Headline result	Hallucinations "occurring between 58% of the time with ChatGPT 4 and 88% with Llama 2"	Each legal tool "hallucinate[s] between 17% and 33% of the time"
Other findings	LLMs "often fail to correct a user's incorrect legal assumptions"; they "cannot always predict, or do not always know, when they are producing legal hallucinations"	Lexis+ AI answered 65% of queries accurately; Westlaw AI-Assisted Research was accurate 42% of the time "but hallucinates nearly twice as often as the other legal tools"; Ask Practical Law AI gave incomplete answers on more than 60% of queries
Publication	*Journal of Legal Analysis* 16, no. 1 (2024): 64-93	arXiv preprint, version of 30 May 2024

Sources: Dahl et al.; Magesh et al..

The definitions matter more than the percentages. Magesh et al. define a hallucination as "a response that contains either incorrect information or a false assertion that a source supports a proposition". They then distinguish a *misgrounded* response, one where "key factual propositions are cited but misinterpret the source or reference an inapplicable source", from an *ungrounded* one, where propositions are not cited at all. A misgrounded answer counts as a hallucination even if every cited case exists.

That distinction is the key to controls. *Mata* involved citations to cases that did not exist. The more common failure in tools that retrieve real documents is a real case cited for a proposition it does not support. A control that only checks existence will catch the first and miss the second.

Why plausible citations are the dangerous failure

A citation such as "Varghese v. China Southern Airlines Co., Ltd., 925 F.3d 1339 (11th Cir. 2019)" has the form of a real citation: party names, reporter, volume, page, court and year. Form is what a reviewer skims. A citation that looks wrong is caught; a citation that looks right is trusted.

Three properties make plausible fabrications hard to catch by reading:

- **The format is learnable.** Reporter abbreviations and court names follow fixed patterns that a language model reproduces well.
- **Partial truth.** A fabricated case may name a real court and real judges, as the "Varghese" opinion did.
- **Self-confirmation.** Asking the generating system whether its output is real, as in *Mata*, asks it to generate more text of the same kind. The Dahl study's finding that models "cannot always predict ... when they are producing legal hallucinations" is the empirical counterpart.

Control 1: resolve every citation to a retained primary source

Rule: no citation enters a filing unless it has been resolved, by a person or by a system independent of the drafting tool, to the text of the cited authority in a primary or authoritative database, and a copy of that text, or a stable reference to it, has been retained with the draft.

Resolution means more than finding a case with a similar name. It means matching the citation's reporter, volume, page, court and year to a single document, and retaining that document. A citation that cannot be resolved is not a "possibly valid" citation; it is unresolved and must be removed or fixed.

Control 2: passage-level support check

Rule: for every proposition supported by a citation, the reviewer identifies the passage in the retained authority that supports it, and records the passage.

This is the control that catches misgrounding. It converts "the case exists" into "this sentence of this case says this". The site's method note on citing at the sentence describes the granularity; the claim-level citation test shows how to evaluate a system against it. In practice, a pin cite to a page is the minimum; recording the quoted passage is better, because it lets a second reviewer check support without re-reading the opinion.

Control 3: a declared "not found"

Rule: the research workflow must have an explicit, acceptable outcome for "no supporting authority found", and that outcome must be recorded rather than filled.

In *Mata*, the firm's research tool did not return federal authority, and the gap was filled by a tool that generated authority. A workflow in which "nothing found" is an acceptable answer, recorded and escalated, removes the incentive to fill the gap. The site's method note on declaring unknowns sets out the general practice. Magesh et al. treat refusals as a separate outcome from hallucinations for the same reason: an incomplete answer is a different, and more manageable, failure than a fabricated one.

A worked example

The following brief is fictional. An associate drafts a two-page letter brief arguing that a contractual limitation period bars a claim. The draft, prepared with an AI drafting assistant, contains eight citations.

Citation	Control 1: resolved?	Control 2: passage supports proposition?	Outcome
1	Yes	Yes, pin cite and quote recorded	Keep
2	Yes	Yes	Keep
3	Yes	No: the passage concerns a statutory, not contractual, period	Misgrounded; replace or drop proposition
4	Yes	Partly: holding is narrower than stated	Revise the sentence

Citation	Control 1: resolved?	Control 2: passage supports proposition?	Outcome
5	No: reporter and page resolve to a different case	n/a	Unresolved; remove
6	Yes	Yes	Keep
7	No: no case under that name in the court and year	n/a	Treated as fabricated; remove and report
8	Yes	Yes	Keep

Five of eight citations survive unchanged. Two would have passed an existence check and failed only at the passage check. One, citation 5, would have passed a casual name search because a case with a similar name exists. The record of this review, with retained copies and passages, is also the firm's evidence if the filing is later questioned.

Court rules and disclosure

Courts responded to *Mata* with standing orders and, in some districts, local rules. They vary, and a lawyer must check the rules of each court. As one example, the Northern District of Texas page for Judge Brantley Starr states that he "no longer requires a certification on the use of artificial intelligence in light of the Court's adoption of Local Civil Rule 7.2(f) and Local Criminal Rule 47.2(e), requiring the disclosure of the use of generative AI in documents that lawyers prepared using generative AI" (N.D. Tex., Judge Starr).

Bar ethics guidance points the same way. The American Bar Association's Formal Opinion 512 (29 July 2024) is its first on generative AI; the ABA's own summary says the model rules on "competency, informed consent, confidentiality and fees principally apply", and that lawyers must "fully consider their applicable ethical obligations", including competent representation, protecting client information, communicating with clients and charging reasonable fees. Its fee example allows billing for the time "to review the resulting draft for accuracy and completeness" (ABA news release, 29 July 2024). This article relies on that summary, not on the full opinion text.

Disclosure rules and the three controls address different things. Disclosure tells the court a tool was used. The controls determine whether the citations are true. A filing can comply with a disclosure rule and still contain a misgrounded citation; Rule 11, which the court applied in *Mata*, concerns the accuracy of the filing regardless of how it was drafted.

What the research says about evaluation

The research community distinguishes between checking that a statement is supported and checking that it is correct. The Attributable to Identified Sources (AIS) framework of Rashkin et al. assesses "whether model-generated statements are supported by underlying sources". The ALCE benchmark of Gao et al. automates citation evaluation and reports that on one dataset "even the best models lack complete citation support 50% of the time". Both frameworks presuppose what Control 1 provides: a retrieved, identified source against which support can be judged.

Checklist

1. Is every citation resolved to a primary or authoritative source independent of the drafting tool?
2. Is a copy or stable reference of each cited authority retained with the draft?
3. Does every cited proposition have a recorded supporting passage or pin cite?
4. Is "no authority found" an accepted, recorded outcome rather than a gap to fill?
5. Are citations that fail resolution treated as fabricated until proven otherwise?
6. Is the AI tool ever used to verify its own output? (It should not be.)
7. Has the court's current local rule or standing order on AI disclosure been checked?
8. When opposing counsel or the court questions a citation, is the first step to re-run Controls 1 and 2, not to defend the citation?

Summary

Mata v. Avianca is a case about verification, not about a tool. The sanctions opinion states that using a reliable AI tool is not inherently improper, and locates the failure in the gatekeeping duty. The empirical studies show that fabrication and misgrounding persist even in retrieval-based legal tools. Three controls address both: resolve every citation to a retained source, check support at the passage, and allow "not found". The site's legal sector note discusses the same requirement for the facts a legal team relies on.

Sources

1. *Mata v. Avianca, Inc.*, No. 22-cv-1461 (PKC) (S.D.N.Y.), Opinion and Order on Sanctions, ECF 54, 22 June 2023 (RECAP copy). https://archive.org/download/gov.uscourts.nysd.575368/gov.uscourts.nysd.575368.54.0.pdf
2. *Mata v. Avianca, Inc.*, docket, CourtListener. https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc/
3. Dahl, Magesh, Suzgun and Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, Journal of Legal Analysis 16(1) (2024): 64-93; arXiv:2401.01301. https://arxiv.org/abs/2401.01301
4. Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, Journal of Empirical Legal Studies 22(2) (2025): 216-242, DOI 10.1111/jels.12413; arXiv:2405.20362 (version of 30 May 2024; same 17% to 33% headline). https://arxiv.org/abs/2405.20362
5. Rashkin et al., *Measuring Attribution in Natural Language Generation Models*, arXiv:2112.12870. https://arxiv.org/abs/2112.12870
6. Gao et al., *Enabling Large Language Models to Generate Text with Citations*, EMNLP 2023, arXiv:2305.14627. https://arxiv.org/abs/2305.14627
7. U.S. District Court for the Northern District of Texas, Judge Brantley Starr, judge-specific requirements. https://www.txnd.uscourts.gov/judge/judge-brantley-starr
8. American Bar Association, "ABA issues first ethics guidance on a lawyer's use of AI tools" (Formal Opinion 512), news release, 29 July 2024. https://www.americanbar.org/news/abanews/aba-news-archives/2024/07/aba-issues-first-ethics-guidance-ai-tools/

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://archive.org/download/gov.uscourts.nysd.575368/gov.uscourts.nysd.575368.54.0.pdf
2. https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc
3. https://arxiv.org/abs/2401.01301
4. https://arxiv.org/abs/2405.20362
5. https://www.txnd.uscourts.gov/judge/judge-brantley-starr
6. https://www.americanbar.org/news/abanews/aba-news-archives/2024/07/aba-issues-first-ethics-guidance-ai-tools
7. https://arxiv.org/abs/2112.12870
8. https://arxiv.org/abs/2305.14627

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.