It Has Sources. Can You Trust It Now?

Part 4 of a series on AI hallucinations

By Dr Peter Fields

Draft: 17 November 2026

First posted: [to be completed on publication]

THE 2-MINUTE SUMMARY

Retrieval gives AI access to material it can use to answer. A search may find a relevant clause while missing the exception or amendment that changes the answer. Even when the right passage is available, a summary can lose a qualification that matters.[1] [2]

A working link tells you where to look. It takes a further check to see whether the passage supports the whole claim. In a historical study of US legal research tools, specialist systems produced fewer hallucinations than the general-purpose comparator, but genuine citations could still accompany wrong answers.[5]

Ask four questions. Does the source exist? Does it support the claim? Does it apply here? Have the necessary qualifications and contrary sources been considered?

Several references may trace back to the same account. For a consequential claim, follow the evidence back to the original and check what it establishes.

The answer comes with six references. The links work. The documents exist.

That is progress. It is also where a different kind of checking begins.

The most conspicuous legal hallucinations involve invented cases. Once we have seen a few, the obvious precaution is to check that the case exists. Yet an answer can cite a genuine judgment and misstate what it decided. It can reproduce an old policy accurately while overlooking its replacement. It can attach a sound source to a sentence that goes considerably further than the evidence.

Sources make an answer easier to investigate. Whether they justify believing it is another question.

What retrieval actually changes

Part 3 introduced retrieval-augmented generation, usually shortened to RAG. The basic arrangement is straightforward: a system searches an external collection, supplies selected material to a language model, and asks it to answer using that material. The original RAG research combined retrieved passages with knowledge held in the model's parameters. It improved factual generation on the tasks tested.[1]

This matters. A model no longer has to rely entirely on what it learnt during training. It can consult information published later, private documents or specialist collections.

But the model still has work to do. It must identify the relevant evidence, understand it, combine it appropriately and keep its answer within what that evidence supports.

Retrieval changes the information available. It does not certify the conclusions drawn from it.

First, find the right material

Imagine asking an assistant whether a company can terminate a supply agreement immediately. This is a hypothetical example.

It retrieves a clause permitting termination for material breach. The answer says that immediate termination is available.

Elsewhere in the agreement, another clause requires written notice and an opportunity to remedy the breach. An amendment extends the time allowed to remedy the breach from 14 to 30 days. A schedule modifies the termination provision for this particular product.

Every passage quoted in the answer might be genuine. The conclusion could still be wrong because the search failed to assemble the material needed to answer the question.

This is a useful distinction between finding relevant text and finding sufficient evidence. A passage can be highly relevant without being sufficient.

The same problem arises outside law. An expenses policy may contain a general entitlement and a separate exception. A scientific article may report a result in the abstract and qualify it in the methods or limitations. A product manual may describe a feature available only on one model.

The evidence that defeats an attractive answer may be in a different place from the evidence that first suggests it.

Having the passage does not ensure it is used properly

Retrieval is only one possible point of failure. A model can receive the necessary material and still mishandle it.

RAGTruth, a research corpus published in 2024, contains nearly 18,000 generated responses, manually annotated for departures from supplied material. The researchers recorded unsupported additions and contradictions, including subtle changes of meaning. Access to reference material did not reliably prevent those errors.[2]

Return to the hypothetical contract. Suppose the assistant receives the complete termination clause, including the notice requirement, but summarises it as an unconditional right to terminate immediately. The evidence was available. The answer failed to preserve it.

Small words can carry large consequences: “may”, “must”, “unless”, “only”, “normally”. A summary that drops one can change the rule while retaining most of the original vocabulary.

This is why a reassuring quotation does not settle the matter. The quotation may be accurate while the sentence introducing it draws an unsupported conclusion.

More documents are not automatically the answer

One response is to give the system everything. Upload the entire agreement, every amendment and the correspondence. Surely the missing qualification will then be somewhere in the material?

Perhaps. But availability and effective use are different.

The research known as *Lost in the Middle* tested how models used information in long inputs. Performance often deteriorated when relevant material appeared in the middle, rather than near the

beginning or end. Those experiments involved earlier models and specified tasks. They do not establish the same performance pattern for every current system.[3]

They do establish a reason to test an assumption: the capacity to accept a long document is not, by itself, evidence that every part will be used reliably.

For the person commissioning the work, “the file was uploaded” and “the relevant provisions were checked together” are different statements. The second needs evidence of its own.

A real source may not support the claim

A citation answers at least two questions. What document is being identified? What does that document establish?

Those questions need separate checks.

A real paper may concern a different population. A genuine judgment may discuss an argument that the court rejected. An authentic policy may apply to employees in a different country.

A reference at the end of a sentence may support only half of it. Consider this hypothetical sentence:

The intervention improved performance and is suitable for all schoolchildren.

A study might support the first proposition in a particular experiment. It might say nothing about universal suitability. The reference exists, the topic matches and the opening proposition is defensible. The final claim remains unsupported.

Research on generating text with citations treats the quality of support as something to evaluate separately from fluency and answer correctness. The ALCE benchmark distinguishes whether statements receive adequate citation support from whether citations are relevant to those statements.[4]

That distinction is useful well beyond benchmarking. Counting references tells us little about how much of an answer they actually substantiate.

What the legal research study showed

A study by Magesh and colleagues tested US legal research tools using queries run between March and May 2024. Published in the *Journal of Empirical Legal Studies* in 2025, it reported hallucination rates between 17% and 33% across the three specialist tools evaluated.[5]

The study examined claims made in marketing these tools. The authors quote LexisNexis as promising “100% hallucination-free linked legal citations”. The researchers counted false statements and false claims about what a source supported. A link could lead to a genuine judgment and still fail that second test.[5]

All three specialist tools produced fewer hallucinations than the general-purpose GPT-4 comparator in this evaluation. The study also reported incomplete responses separately. As Part 3 showed, counting fewer wrong answers tells us only part of the story.[5]

These were particular tools, versions, queries and assessment criteria. The results are not current product rankings, general error rates for legal practice, or findings about English legal research.

Their enduring lesson is narrower: access to a specialist collection did not eliminate errors in the systems tested. The relationship between the cited authority and the answer still required scrutiny.

Faithful to a source, wrong about the world

There is another distinction that can get lost in discussions of hallucination.

An assistant can reproduce a source faithfully and still give an unreliable answer because the source itself is wrong, outdated or inappropriate.

Suppose an internal handbook says that a particular approval process applies. The assistant accurately explains that process. Unknown to it, a later policy has replaced the handbook.

The output may be faithful to the retrieved text. It is nevertheless an unsafe account of the present position.

Conversely, an answer might state something true but attribute it to a document that does not establish it. Its conclusion could survive, while its claimed evidential basis fails.

Part 1 introduced the distinction between faithfulness to a source and factuality. The first question is whether the answer represents the material accurately. The second is whether the claim is true. Even an accurate account of a source can leave that second question open.[6]

GEEK BOX: Four checks behind a citation

A useful way to assess a cited claim is to ask four questions:

1. **Existence:** does the identified document exist, and have we accessed the correct one?

2. **Support:** does the relevant passage support the whole proposition being advanced?

3. **Applicability:** is the source appropriate to this date, jurisdiction, population or set of circumstances?

4. **Coverage:** have other necessary provisions, qualifications or conflicting sources been considered?

An answer can pass the first check and fail any of the others.

These are an editorial framework for checking, rather than a validated scoring system. They explain why a working hyperlink is the beginning of verification.

What if the source has been manipulated?

So far, the examples have involved ordinary research failures. Retrieved material can also be deliberately misleading.

PoisonedRAG, presented at USENIX Security in 2025, demonstrated attacks in which malicious text inserted into a knowledge collection steered tested systems towards answers chosen by the attacker.[7]

This was an experimental security finding under specified conditions. It does not mean that any web page can reliably control any assistant.

It does show why connecting a model to external information introduces a question about that information's integrity. A system may follow a source closely and thereby transmit the attacker's falsehood.

That is a different mechanism from spontaneously inventing a fact. From the reader's perspective, however, the result can look similar: a confident, referenced answer that should not be relied upon.

Do several sources provide independent confirmation?

They can, provided they are independent and actually address the claim.

Three websites repeating the same press release are three locations for one account. A commentary quoting another commentary may add no independent evidence. A search result that repeats an AI-generated claim does not become corroboration simply because it appears outside the original conversation.

In *Campos & Munoz*, discussed in Part 2, fictitious authorities appeared in a signed family court order. A later reader could find them in a genuine court document, but their inclusion would supply no independent confirmation that the cases existed.[8]

This connects with the problem examined in Part 2. Asking for evidence after a doubtful answer is useful only if the next step obtains evidence capable of testing it. Another confident repetition does not do that.

The practical question is where the information originated and whether the sources bring independent grounds for accepting it.

For an important claim, following the chain back to the underlying judgment, dataset, policy or experiment is often more valuable than accumulating further summaries.

What sources genuinely give us

The answer is substantial: access to evidence, a means of updating information, and a route by which the reader can examine the basis of a claim.

Used well, retrieval can reduce dependence on uncertain recall. Precise citations can make mistakes easier to identify. A passage, page number and explanation of the link between source and conclusion can save considerable checking time.

The danger is treating that opportunity to verify as if verification has already occurred.

My conclusion is that confidence should depend on the relationship between the claim and the evidence. For consequential work, the system should show the relevant passage, preserve qualifications, distinguish quotation from inference and identify gaps it cannot resolve.

An answer that says “these documents do not establish the point” may be the most useful result of a properly conducted search.

The presence of sources is a reason to inspect the answer more efficiently. Whether to rely on it depends on what that inspection establishes.

Part 5 turns to the claims used to compare systems: what does a “hallucination rate” measure, and what would it take to show that the problem has been solved?

Sources and further reading

1. Lewis et al. (Facebook AI Research, UCL and NYU), *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, NeurIPS (2020), particularly §§2 and 4. Paper.

2. Niu et al., *RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models*, ACL (2024), pp 10862–10878, particularly §3.1. Paper.

3. Liu et al., *Lost in the Middle: How Language Models Use Long Contexts*, TACL 12 (2024), pp 157–173; first released in 2023. Historical models and controlled tasks. Paper.

4. Gao, Yen, Yu and Chen, *Enabling Large Language Models to Generate Text with Citations*, EMNLP (2023), particularly §3.3. Paper.

5. Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, Journal of Empirical Legal Studies 22(2) (2025), pp 216–242, particularly §§1 and 4–6, including §4.3 on the treatment of linked citations. Queries run in 2024. Full text.

6. Maynez et al., *On Faithfulness and Factuality in Abstractive Summarization*, ACL (2020), particularly §2. Paper.

7. Zou et al., *PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models*, USENIX Security (2025), pp 3827–3844. Paper and proceedings.

8. *In re Domestic Partnership of Campos & Munoz*, No D085584 (Cal Ct App, 5 March 2026, subsequently modified), part I. Judgment with modification.