Can You Use AI Without Being Misled?

Part 6 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

Before asking AI to do consequential work, ask yourself one question. If this answer were wrong, how would I find out? Start with a task you can check and decide what evidence you will need before you read the answer.

Ask the assistant to show the passage behind each material claim and identify what it inferred. Then open the source. The Divisional Court emphasised lawyers' duty to check AI-assisted legal research against authoritative sources. Assign that check to someone with the time, access and expertise to do it.[1]

If you find a material mistake, revisit every conclusion that depends on it. Judge the time saved by the finished job, including the work of checking and correcting it.

An assistant produces a convincing answer in seconds. Checking it takes twenty minutes.

Has it saved any time?

Sometimes, yes. Finding the relevant passages, organising a difficult question and preparing a usable first draft may have taken much longer without it. Sometimes, no. An impressive answer can create more work than it removes, particularly when its errors are scattered through otherwise useful material.

That is the practical question at the end of this series. We know that AI can produce plausible falsehoods. We also know that better training, retrieval and checking can reduce them, without providing a general guarantee of truth.

How should we work with that combination?

My answer is to choose tasks with a workable route to verification, keep evidence visible and decide who will check what before relying on the result. The working method below is my synthesis of the evidence discussed throughout the series. It is not a protocol proven to eliminate hallucinations.

Choose work you can check

The first decision comes before the prompt.

An assistant asked to shorten a letter can be checked against the original. An assistant asked to invent ten advertising slogans can be judged against the brief. An assistant asked to identify a decisive but obscure legal authority may produce an answer whose defects require specialist research to discover.

These are different propositions for the user, even if the same interface handles all three.

A useful starting point is to ask: if this answer were wrong, how would I find out?

If the answer is “I would recognise it immediately”, the checking burden may be modest. If it is “I would need expertise I do not possess”, the system may help you explore the subject, but its answer cannot supply the missing assurance.

Supplied documents make some checks easier. They do not make the task automatically safe. As Part 4 explained, an accurate quotation can sit beside an unsupported conclusion, and a summary can omit the qualification that changes everything.

Choose the work partly by how readily its success can be assessed.

Define the task before asking for the answer

A broad request invites the assistant to make choices about scope, assumptions and completeness.

Suppose you ask, “Can we terminate this agreement?” The answer may depend on which version applies, what happened, whether notices were served and what jurisdiction governs the dispute.

A more controlled first task would identify the termination provisions, extract their conditions, locate amendments and list the facts still needed. The legal conclusion can then be considered against that record.

This does not remove the need for judgement. It makes the steps open to inspection.

The same approach works with less specialised material. Before asking for a conclusion about a supplier’s performance, ask which deliveries were late, what records establish the dates and which dates remain disputed. Before asking for an account of a scientific finding, establish the study population, comparison, outcome and limitations.

A well-defined intermediate task can be more useful than a prematurely complete answer.

Make the evidence visible

For factual work, I would ask the assistant to separate what the material states from what it infers.

A document may record that a meeting occurred. The assistant might infer that a decision was made there. Those propositions should not quietly merge.

Likewise, a witness may allege that a payment was promised. A chronology should identify the allegation and its source, rather than convert it into an agreed fact.

Precise references help: a document name, page or paragraph, and the passage supporting the claim. They give the reviewer somewhere to go. A list of links at the end of a polished report leaves considerably more work to do.

But asking for a passage is an instruction to produce something checkable. It does not establish that the passage has been reproduced accurately, or that it supports the conclusion. Open the source and apply Part 4’s four checks. Does it exist? Does it support the claim? Does it apply here? What else might change the answer?

GEEK BOX: A prompt that exposes gaps

The following is a working instruction, not a guarantee of accuracy:

Work from the supplied documents. For each material factual claim, identify the document and page or paragraph supporting it. Distinguish direct statements in the documents from your inferences. Preserve qualifications and conflicting accounts. If the available material does not establish a point, say so and identify what is missing. Do not fill gaps with plausible details. Before drafting the conclusion, list any unresolved issues that could change it.

This prompt makes compliance easier to inspect. It cannot ensure compliance.

Where external research is required, define that separately, including the relevant date and jurisdiction. Do not let an answer silently move between supplied evidence and outside material.

Treat instructions to be careful as instructions

There is a temptation to search for the prompt that makes hallucinations disappear: "be accurate", "do not invent", "answer only if certain".

These instructions can express what we want. They are not evidence that we have received it.

In its 2025 judgment in *Ayinde* and *Al-Haroun*, the Divisional Court dealt with two cases in which lawyers had put fictitious authorities before the court. Five cases cited in *Ayinde* did not exist, and the source of those citations remained unresolved. In *Al-Haroun*, 18 of the 45 citations in a schedule referred to cases that did not exist.[1]

The court concluded that asking the tool to decline when unsure or to identify weaknesses in a case did not materially reduce the risks it was discussing. At [87], it put the point directly. "The critical safeguard is to check any output by reference to an authoritative source."[1]

Check claims through a different route

Part 3 discussed Chain-of-Verification, in which a draft is broken into checking questions that are answered separately before revision. The experiments reduced hallucination on the tasks tested. They did not turn the model's own answers into external corroboration.[2]

The useful principle is to change the checking task.

Instead of asking whether a proposed notice period is correct, inspect the operative clause and calculate the period from the relevant dates. Instead of asking whether an article supports a conclusion, read its results and limitations. Instead of asking whether a quotation is authentic, find it in the original document.

Another assistant can help identify weaknesses or suggest searches. In *Kenney*, discussed in Part 2, counsel said her staff had used Google, ChatGPT and Microsoft Copilot to cross-check citations. The appellate court criticised the subsequent submission of AI-generated case summaries to explain the fictitious authorities.[3]

Several tools can repeat the same mistake. For a case citation, the check is to find and read the judgment.

Do not confuse agreement with correction

A challenge can produce a more accurate response. It can also produce an unwarranted apology and a new mistake.

Research on sycophancy found that historical assistants could abandon correct answers when challenged and accommodate users' mistaken beliefs. These were particular models and experiments, not a measured failure rate for every present system.[4]

The implication for practice is straightforward: neither confidence nor capitulation settles the point.

"Are you sure?" may be a useful conversational interruption. The next useful question is what evidence decides the issue.

Equally, asking the assistant to be adversarial does not make its objections sound. A list of invented difficulties is no better than invented support. Test criticism by the same standard as endorsement.

The aim is a conclusion that survives scrutiny, whether or not it is the conclusion you wanted.

Give human review a specific job

"Human in the loop" sounds reassuring. It leaves unanswered what the human is expected to do.

Reading for plausibility is different from checking a quotation. Checking a quotation is different from establishing whether the authority remains applicable. Reviewing five sample entries is different from reviewing every material assertion.

For a consequential document, I would allocate the checks explicitly. Who verifies names and dates? Who checks calculations? Who reads the cited authorities? Who decides whether the conclusion follows?

The reviewer needs the relevant competence, access to the evidence and enough time to use both. Merely placing a person between the model and the recipient supplies none of those things.

In *Al-Haroun*, the solicitor relied on legal research supplied by his lay client. The Divisional Court described that reliance as "extraordinary". Accuracy of the citations and quotations put before the court remained the lawyer's professional responsibility. The judgment also makes clear that lawyers' duties extend to AI-assisted research done by others on their behalf.[1]

Watch what happens to your own confidence

The assistant's fluency is only one part of the problem. The other is the way we respond to it.

A 2025 study surveyed 319 knowledge workers about their use of generative AI. Greater confidence in AI was associated with less reported critical thinking; greater task-specific self-confidence was associated with more. The study concerned self-reported behaviour and associations. It did not establish that AI caused a decline in thinking ability.[5]

Nevertheless, it identifies a practical tension. Feeling confident about the tool may coincide with doing less of the work needed to check it.

My proposed response is to decide the checking standard before reading the answer. If a deadline, quotation or legal proposition requires verification, its smooth presentation should not reduce that requirement.

An answer can be immediately useful without being ready to rely upon. Keeping that distinction visible is part of the user's work.

When a material error appears, reopen the conclusion

Finding an error should change the review, not merely the sentence containing it.

Suppose a report relies on the wrong contract version. Correcting its title is insufficient. The notice analysis, payment provisions and recommendations may all need to be reconsidered.

If a case cannot be found, stop using it while investigating. Seek the authority through an appropriate independent source. Do not invite the assistant to manufacture an increasingly elaborate account of where it supposedly appeared.

If verification remains unsuccessful, record that the proposition has not been established. That is different from proving it false, but it is sufficient reason not to present it as verified.

Where inaccurate material has already been communicated, consider the corrections required in that setting. Preserve the relevant record and identify conclusions affected by the mistake.

This is the point at which the human can interrupt the escalation described in Part 2. The need to defend a finished document must not dictate what counts as evidence for it.

Measure the completed job

The time taken to generate an answer is only part of the cost.

Include the work needed to prepare the material, check the output, repair errors and complete the task. Consider omissions as well as false statements. An answer containing nothing invented may still fail because it leaves out the decisive issue.

For repeated work, a small record of recurring mistakes can guide changes: narrower tasks, clearer sources, different tools or additional review. Test those changes on work representative of the actual task.

Sometimes the result will justify extensive use. Sometimes a conventional search, spreadsheet or direct reading will be quicker and more dependable.

The objective is better completed work.

What it means to rely on AI

The question running through this series has been whether we can rely on a system that sometimes makes things up.

Reliance need not be an all-or-nothing decision about the assistant. It can be a specific decision about a particular contribution: a passage located, a draft reorganised, a calculation checked or an alternative explanation tested.

The more consequential the conclusion, the more clearly we need to understand its evidential basis and the work still required from us.

AI can help us find, organise and examine information. Used carelessly, it can also make an unsupported conclusion easier to produce and harder to notice.

The practical standard I would adopt is this: know what the assistant contributed, know how the material claims were checked, and be able to explain why the final conclusion is justified.

Think back to *Mata v Avianca*, where this series began. The question to ask before filing was where the judgments themselves could be found. ChatGPT's assurances could never supply that evidence.[6]

Sources and further reading

1. *R (Ayinde) v London Borough of Haringey; Al-Haroun v Qatar National Bank* [2025] EWHC 1383 (Admin), particularly [7]–[9], [46]–[48], [74], [79]–[81] and [87]. Judgment.

2. Dhuliawala et al. (Meta AI and ETH Zürich), *Chain-of-Verification Reduces Hallucination in Large Language Models*, Findings of ACL (2024), particularly §§3–4. Paper.

3. *In re Sanctions Order of Kenney, Kerry*, No 25-C-389 (La App 5 Cir, 23 October 2025), printed pp 4–5 and 18–20. Judgment.

4. Sharma et al. (Anthropic-led), *Towards Understanding Sycophancy in Language Models*, ICLR 2024 (first circulated 2023), particularly §3.2. Paper.

5. Lee et al. (Microsoft Research), *The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers*, CHI (2025), article 1121, pp 1–22. Research publication.

6. *Mata v Avianca, Inc.*, 678 F Supp 3d 443 (SDNY, 22 June 2023), particularly Findings of Fact [45]–[47]. Sanctions order.