What Does a Hallucination Rate Actually Tell You?

Part 5 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

When you see a hallucination rate, ask what was tested and what was counted. A score based on summaries of supplied documents tells you about a different job from a score based on answers to legal research questions. The scoring rules matter too.[1] [2]

Look for correct answers, wrong answers and questions the system declined. A tool can make fewer mistakes by answering fewer questions. And one wrong deadline can matter more than dozens of facts the tool got right.

A mathematical lower bound can help explain some errors under specific assumptions about training and the model. The same research gives grounds for thinking that recurring publication references can be handled better. A system with a well-defined remit can answer within it and decline the rest.[6] [7]

An AI system is said to hallucinate on only 1% of occasions.

That sounds reassuring. But 1% of what? Answers, sentences, factual claims, citations, or questions the system agreed to attempt? Which questions? Under what conditions?

Until we know, the number tells us surprisingly little about whether to trust the next answer.

Vectara's leaderboard on 22 October 2026 reported hallucination rates as low as 1.8% for models summarising supplied documents, using its own AI grader.[1] The study discussed in Part 4 reported rates of 17% to 33% for three specialist tools tested on US legal research questions in 2024. Its researchers scored the answers by hand.[2] Both sets of figures carry the same label. Each describes performance on a particular task under its own scoring rules.

The earlier articles examined why hallucinations arise and what can reduce them. This article turns to the evidence used to measure progress. It also addresses a claim that can sound equally conclusive: that mathematics proves hallucinations can never be eliminated.

Both subjects require attention to what has actually been established.

Start with the question being tested

There is no reason to expect an assistant to perform identically when answering an obscure historical question, summarising a supplied letter and researching a disputed legal proposition.

These tasks place different demands on it. The first may test stored knowledge. The second tests fidelity to a source. The third may require retrieval, interpretation, comparison and attention to jurisdiction and date.

SimpleQA, released in 2024, deliberately narrowed the problem. It comprised 4,326 short factual questions intended to have single, verifiable answers. Questions were selected to challenge the models used during collection. This made factual performance easier to assess, while leaving open whether gains would transfer to longer answers.[3]

A difficult test can be useful precisely because it exposes weaknesses. Its results do not automatically describe ordinary user traffic.

An examination designed to catch out well-prepared candidates is not a random sample of the questions they encounter at work.

The denominator changes the meaning

Part 3 distinguished correcting an answer from withholding it. Here is what that distinction does to the headline rate. In the table below, Assistant B gives a wrong answer to 5% of all questions it receives. Among the 65 it attempts, the figure is 7.7%.

To see the trade-off, ask for the counts of correct answers, wrong answers and abstentions, together with the error rate among attempted answers.

GEEK BOX: The same experiment, different headlines

These figures are invented to illustrate the arithmetic. They are not results for any commercial product.

Outcome or measure	Assistant A	Assistant B
Questions received	100	100
Correct answers	80	60
Incorrect answers	20	5
Questions declined	0	35
Incorrect answers as a proportion of all questions	20%	5%
Incorrect answers as a proportion of attempted answers	20%	7.7%
Proportion of questions attempted	100%	65%

B produces 75% fewer wrong answers in this example. It also produces 25% fewer correct answers. Neither fact should disappear from the comparison.

The preference depends on the task and the relative costs of a wrong answer and an unanswered question.

An answer is not a single fact

A short response may contain one material assertion. A detailed report may contain dozens.

FActScore, introduced in 2023, addressed this by breaking generated text into individual factual claims and assessing what proportion was supported by a specified knowledge source. That offers a more detailed view than marking an entire passage simply right or wrong.[4]

But a claim-level figure and an answer-level figure remain different measures.

Imagine 100 answers containing 20 claims each: 2,000 claims in total. Suppose 20 claims are wrong, giving an overall claim error rate of 1%.

If those errors occur in 20 separate answers, one in five answers contains an error. If all 20 occur in one answer, only one in 100 answers contains an error.

The claim error rate is identical. The distribution of affected answers is very different.

This illustration makes no assumption that errors occur independently. That matters because real mistakes can cluster: one wrong identification may contaminate several later claims.

A low average can conceal the mistake that matters

Counting claims also leaves a question about their importance.

A report might correctly identify the parties, procedural history, hearing date and background facts, then get the decisive limitation date wrong. A high proportion of correct statements would offer little comfort to the person relying on that date.

Conversely, an otherwise useful summary might contain a minor error in an incidental biographical detail. Both are errors. Their consequences differ.

For consequential work, I would want evaluation to identify material mistakes separately. Does an error change the recommended action, the sum payable, the deadline or the conclusion?

An aggregate score cannot make that judgement for every intended use. The person choosing the system must connect the measurement to the work.

Who checks the checker?

A benchmark needs reference answers and a method of judging responses. Neither should be assumed infallible.

SimpleQA used an automated model to classify answers against reference answers. The authors estimated that their benchmark had an error rate of around 3%, assuming the automated grader produced no false positives.[3] Later, Google researchers produced SimpleQA Verified, a revised 1,000-question benchmark addressing problems including reference-answer errors, redundant questions and uneven topic coverage.[5]

That is evidence of evaluation being examined and improved. It does not make the original research worthless, or establish that every revision is beyond criticism.

It means that the measuring instrument has its own limitations.

A claim can be marked unsupported because supporting evidence was missed. An ambiguous question can admit more than one defensible answer. A judge can misunderstand a qualification.

Human review has costs and fallibility too. The appropriate response is to examine how grading was checked, how disagreements were handled and whether reported differences are large enough to justify the conclusion drawn.

Compare the systems actually tested

An assistant with browsing enabled is doing a different job from a model answering from stored knowledge. A system allowed several searches, repeated attempts and a verification stage has different resources from one required to answer immediately.

Those differences may be entirely appropriate. We often care about the performance of the complete service we will use.

But the comparison must say what was provided. Otherwise, an improvement in the surrounding workflow may be attributed to the underlying model, or a restricted test may be mistaken for the capability of the complete product.

The model version, test date, instructions, available tools and checking procedure all help define the result.

The same applies to a claimed reduction. A fall from 20% to 10% is a 50% relative reduction and a fall of ten percentage points. A headline announcing “50% fewer hallucinations” leaves the starting point unstated.

Does mathematics prove hallucinations are inevitable?

Some theoretical results are valuable explanations of why errors arise. Their assumptions matter as much as their conclusions.

Kalai and Vempala’s *Calibrated Language Models Must Hallucinate* analyses a statistical setting involving certain arbitrary facts and a particular condition on the model’s probability distribution. Under its assumptions, the paper establishes lower bounds on hallucination.[6]

That is a conditional result. It does not establish an unavoidable error rate for every assistant, task or checking process.

The bound is tied to the share of arbitrary facts that occur only once in the training sample. The authors give references to articles and books as an example of information that tends to recur. Their argument gives no statistical reason why pretraining would force a model to invent references of that kind. Other causes remain possible. The authors also explain that features absent from their simplified setting, such as documents containing several facts, can affect whether the bound applies.[6]

In Part 2, calibration meant checking whether stated confidence matched observed accuracy. An assistant saying “90% confident” gives a figure that can be tested across answers. Kalai and Vempala use a different yardstick. They group statements the model assigns similar probabilities and compare those probabilities with how often the statements occur, on average, in the kind of text used for training.[6]

A theorem can explain a source of difficulty without proving that every practical way of avoiding it must fail.

A system can know its limits

The later paper *Why Language Models Hallucinate*, by Kalai and colleagues, explicitly describes a bounded system that answers a fixed collection of questions, performs specified calculations and otherwise declines. Such a system need not fabricate answers.[7]

That is a useful counterexample to the sweeping claim that every AI system must hallucinate.

It does not show that a broad conversational assistant can answer arbitrary questions accurately. Nor does it establish that any particular implementation has perfect data, code or safeguards.

The practical challenge is to retain useful breadth while controlling error.

A system that declines everything avoids factual fabrication but does no useful answering. A carefully restricted system may provide dependable answers within its boundary. An assistant that ranges across almost any subject has a much wider boundary to police.

These are different propositions. Calling hallucinations either “inevitable” or “solved” can obscure the choices between them.

What would count as solving the problem?

We need a specification before we can assess the claim.

“Never invent a case citation” is narrower than “never misstate the law”. “Copy this field exactly from this record” is narrower than “produce a complete and accurate account of this dispute”.

A convincing claim about a restricted task should identify the permitted inputs, required outputs and circumstances in which the system must decline or seek review. It should then provide evidence relevant to that specification.

Zero observed errors in a finite test is evidence about the tested cases. By itself, it is not proof that every future case will be handled correctly. Stronger assurance may come from the way a restricted process is constructed and verified, as well as from testing.

For a broad assistant, “solved” would need to cover an enormous and changing range of questions. A strong score on one benchmark cannot carry that burden.

Equally, an unresolved general problem does not erase useful improvements on narrower tasks.

The figure that matters is the one relevant to your work

My conclusion is that a hallucination rate is useful when it comes with enough information to interpret it: what was tested, what counted as an error, what formed the denominator, how abstentions were treated and how judgements were checked.

Then comes the practical assessment. Which errors matter here? How readily will they be detected? What work remains for the human, and is the resulting process better than the available alternative?

The evidence supports progress, persistent limitations and opportunities to build reliable restricted systems. It does not justify a universal percentage for “how often AI makes things up”, or a blanket mathematical excuse for accepting fabrication.

Part 6 brings the series back to practice: how to use an assistant productively while designing the work around its capacity to be confidently wrong.

Sources and further reading

1. Vectara, *Hallucination Leaderboard*, snapshot dated 22 October 2026. Automated evaluation of summaries of supplied documents. Dated chart. Leaderboard and method.

2. 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, especially §§4–6. Queries run in 2024. Full text.

3. Wei et al. (OpenAI), *Measuring Short-Form Factuality in Large Language Models* (2024), particularly §§2–4. Paper.

4. Min et al., *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*, EMNLP (2023), pp 12076–12100. Paper.

5. Haas et al. (Google DeepMind and Google Research), *SimpleQA Verified: A Reliable Factuality Benchmark to Measure Parametric Knowledge* (2025), particularly §§2–3. Paper.

6. Kalai and Vempala, *Calibrated Language Models Must Hallucinate*, STOC (2024), pp 160–171; preprint first released in 2023. Particularly the introduction, assumptions and limitations. Preprint.

7. Kalai, Nachum, Vempala and Zhang (OpenAI and Georgia Tech), *Why Language Models Hallucinate* (2025), particularly §3.1 and the subsection on base models. Paper.