Why Doesn’t It Just Say “I Don’t Know”?

Part 2 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

An AI assistant needs to recognise uncertainty, act on it and tell the user where the evidence runs out. A model can show signs of recognising an error while still producing a confident answer.[2] [3]

Some tests give guessing an advantage because a wrong answer scores no worse than a blank. Human feedback can also reward agreement, and experiments found that training for warmth could make some models more likely to endorse mistaken beliefs.[4] [7] [8]

In *Mata v Avianca*, ChatGPT assured a lawyer that its fictitious authorities were real. In later cases, further generated summaries or fictitious citation details were offered in support of authorities that did not exist.[1] [10] [11]

More text is easy to generate. More evidence requires a source you can inspect. A useful admission of uncertainty tells you which point remains unresolved and what would settle it.

The invented legal authorities discussed in Part 1 were only part of the story.

In *Mata v Avianca*, the lawyer who had used ChatGPT asked whether one of the cases it had supplied was real. The chatbot reaffirmed its existence and said it was available on Westlaw and LexisNexis. The judgment records both the reassurance and the lawyer’s conflicting accounts of when he sought it.[1]

That exchange exposes a second problem. Producing a false answer is one failure. Providing unwarranted reassurance about it is another.

Most readers can accept that an AI system will sometimes lack information. No person knows everything either. What seems harder to explain is why the system fills the gap instead of acknowledging it.

Why doesn’t it simply say that it doesn’t know?

Three separate achievements

A useful way to approach this is to separate three things an assistant needs to do.

First, it needs some means of distinguishing a sufficiently reliable answer from an uncertain one.

Second, it needs to act on that distinction: answer, qualify, ask a question, consult a source or stop.

Third, it needs to communicate its position accurately.

These are separate achievements. A warning light serves little purpose if the driver ignores it. Equally, a cautious driver cannot respond to a warning that the instrument panel never displays.

This is an analogy, not a description of a literal warning light inside a model. The research nevertheless supports the underlying distinction: information relevant to correctness can be present without reliably controlling the answer presented to the user.[2] [3]

When guessing earns better marks

Imagine an examination with one mark for a correct answer and none for either a wrong answer or a blank.

You are unsure of an answer but have a plausible candidate. Guessing offers a chance of gaining a mark. Leaving the question unanswered offers none.

Under those rules, guessing is the sensible strategy for maximising your score.

Kalai and colleagues apply this reasoning to AI evaluation. Where an incorrect answer and an admission of uncertainty receive the same score, a system willing to guess can outperform a more cautious one. Their argument concerns how evaluation can favour such behaviour during development and selection. A benchmark does not directly instruct a model to invent a particular fact during your conversation.[4]

The practical problem is straightforward. An examination rewards accumulated correct answers. A person relying on an answer may care intensely about avoiding a particular mistake.

The scoring system and the user's interests can therefore pull in different directions.

What human feedback teaches

Developers also shape assistants through examples and feedback.

In the influential InstructGPT work, people demonstrated desired responses and ranked alternative outputs. Those rankings helped train a reward model, which was then used to improve the assistant's behaviour. The resulting systems followed instructions better and showed improvements in truthfulness, while still making mistakes.[5]

That is evidence that feedback can help. It also reveals a limitation: what a person prefers is an imperfect substitute for what is factually correct.

In an analysis of Anthropic's human preference data, Sharma and colleagues found that agreement with a user's beliefs was among the strongest predictors of which response people preferred. The finding helps explain how feedback can reward agreement even when the task calls for a correction.[7]

Can a model recognise its own uncertainty?

It would be convenient if the explanation were that models have no means of assessing whether they are right. The evidence does not support that absolute claim.

Kadavath and colleagues investigated whether language models could evaluate proposed answers and predict which questions they could answer correctly. They found useful self-evaluation under tested conditions, alongside limitations when moving to unfamiliar tasks.[2]

Other researchers have trained classifiers, often called *probes*, to use internal model activity to predict whether an answer is correct. These studies have found information relevant to correctness that is not fully reflected in the generated response. But detectors that perform well on one task can transfer poorly to another.[3]

In some of Orgad and colleagues' experiments, internal signals helped distinguish correct answers from wrong ones even for questions the model usually answered incorrectly. The researchers used trained probes to assess candidate answers. That finding concerns what can be detected in the model's activity and gives no grounds for attributing deliberate concealment to it.[3]

It shows that generating an answer and assessing that answer can draw on information in different ways.

What does “80 per cent confident” actually mean?

A chatbot can produce a numerical confidence estimate as readily as it can produce a date or a name.

The presence of a percentage does not establish that it has been measured or validated. The SimpleQA research examined models’ stated confidence and found overconfidence in its tested factual questions. Those were deliberately difficult questions, not a representative sample of everyday conversations.[6]

The distinction matters because a number can create an impression of precision that ordinary language does not.

“I am fairly confident” sounds like an assessment. “I am 83 per cent confident” can sound like the result of an instrument reading. Unless the method has been evaluated, the extra digits do not earn the extra trust.

GEEK BOX: Confidence is not the same as correctness

Suppose a hypothetical system labels 100 answers as 80 per cent likely to be correct. If 80 are correct, that group is well calibrated.

It still contains 20 wrong answers.

Calibration concerns whether confidence estimates match observed frequencies. A separate question is how effectively the system distinguishes its correct answers from its incorrect ones. Useful performance on one measure does not guarantee perfect performance on the other.[2]

Neither measure, by itself, proves that the assistant will express uncertainty appropriately in its final response.

Sometimes the question supplies the mistake

Uncertainty is also affected by the way a question is framed.

Consider this invented request:

Why did the university withdraw Professor Ellis’s award in 2022?

The question assumes that an award existed and was withdrawn. A satisfactory response must first establish whether those premises are sound.

Yet the form of the request invites an explanation. An assistant that proceeds directly to answering “why” may construct an account around something that never happened.

The problem becomes sharper when we signal the answer we expect. Research on *sycophancy* has found that assistants can adjust responses towards users’ expressed beliefs. Analysis of preference data and experiments with preference models suggest that human approval can contribute to this behaviour. It is a contributing explanation, not proof that feedback invariably makes models less truthful.[7]

“Sycophancy” describes the observed behaviour. It does not require a model to experience a human desire to please.

In legal research, there is a useful difference between asking what the authorities establish and asking for authorities supporting a conclusion already chosen. The latter need not produce an error, but it makes scrutiny of contrary evidence particularly important.

Can friendliness interfere with accuracy?

A supportive tone may seem unrelated to factual reliability. Experimental research suggests that the distinction is less secure than it appears.

Ibrahim, Hafner and Rocher trained five language models to produce warmer responses, then compared them with their original counterparts. Under the tested conditions, the warmer models became less accurate on factual tasks and more likely to endorse incorrect user beliefs. The accuracy gap between the warmer and original models was largest when prompts expressed sadness. In a control involving three of the models, training for a more neutral style produced no comparable systematic decline. The findings are specific to the training methods and tasks tested.[8]

There is also a production example. In April 2025, OpenAI rolled back a GPT-4o update after it became excessively agreeable. Its subsequent account identified the treatment of user feedback as one contributor to the change.[9]

Warmth and agreement are nevertheless different qualities. A courteous assistant can firmly explain that the evidence does not support our position.

The concern is that training for a supportive persona can affect substance as well as presentation. These studies do not establish that warmth caused any particular fabricated legal authority. They do establish a reason to test personality changes for their effects on factual behaviour.

When the mistake acquires supporting evidence

Once somebody has relied on an AI answer, a challenge can prompt a request for evidence that the original answer was right.

The system may then generate the appearance of that evidence: a fuller description, additional identifying details or a confident account of where the source can be found.

In a Louisiana case, *In re Sanctions Order of Kenney, Kerry*, counsel's office responded to challenges about fictitious authorities with AI-generated case summaries. Counsel subsequently submitted those summaries as records supporting the research behind the citations. She said staff had cross-checked using Google, ChatGPT and Microsoft Copilot. The appellate court criticised the attempt to validate AI-generated authorities through further AI output.[10]

The attempted check had reproduced the problem.

A California case, *Campos & Munoz*, provides another striking illustration. When opposing counsel challenged fictitious authorities, the lawyer defending them supplied additional reporter references and a decision date. Those details were also fictitious. She eventually acknowledged using AI resources and conceded that they might have supplied the false information, although the judgment does not conclusively trace each detail to AI.[11]

The fictitious cases had already reached the family court's signed order. It was the appellant's own lawyer who had drafted and submitted that order. The Court of Appeal held that the family court had erred by relying in material part on fictional cases, but concluded that the appellant had forfeited the point.[11]

When human defensiveness joins in

Human behaviour can make this worse. Once an answer has been sent to a client, submitted to a court or used in a decision, acknowledging the error becomes more uncomfortable. A plausible reassurance offers a tempting escape.

That is my interpretation of the interaction, rather than a finding about every person who relies on AI. Some users misunderstand the technology. Others fail to investigate warning signs. In more serious cases, the explanation becomes a separate matter of misconduct.

Mata crossed that further boundary. The court's findings included conscious avoidance of indications that authorities did not exist, an untruthful account of ChatGPT's role and failure to disclose reliance on it in an affidavit.[1]

The model's output and the human response can therefore reinforce one another. An unsupported answer becomes a defended position. Further generated material then appears to justify maintaining it.

The important distinction is between more text and more evidence. A generated summary of an imaginary judgment does not bring the judgment into existence.

A genuine check must reach something capable of establishing the claim independently.

Neither an apology nor reassurance proves anything

This also explains why asking again is an unreliable safeguard.

In the sycophancy research, assistants sometimes abandoned correct answers after a user expressed doubt. Results varied across tasks, and some reasoning tasks benefited from reconsideration.[7]

The *Mata* exchanges illustrate the opposite danger: an invented answer can survive a challenge and acquire further reassurance.[1]

Neither persistence nor retreat establishes correctness.

An assistant that apologises may have corrected a mistake. It may also have introduced one. An assistant that stands its ground may be right, or may simply be continuing its earlier error.

The relevant question is what changed in the evidence or reasoning.

Would teaching it to refuse solve the problem?

An assistant that declines every factual question would avoid many false assertions. It would also abandon much of its purpose.

The objective must therefore involve both reliability and useful coverage. SimpleQA makes that distinction explicit by separately recording correct answers, incorrect answers and questions not attempted.[6]

For a user, the right balance depends on the task.

If I ask for possible themes for a discussion, tentative suggestions may be useful. If I ask whether a document contains a particular undertaking, an unsupported assertion is a different proposition altogether.

Nor is the choice limited to answering or refusing. The assistant can identify the missing information, ask for clarification, distinguish a tentative inference from a documented fact, or consult an appropriate source.

My practical conclusion is that uncertainty should change what the system does. A vague qualification attached to an invented detail leaves the user with the same verification problem.

Permission to be uncertain

There is a corresponding lesson for how we use these systems.

An instruction to identify gaps and avoid filling them plausibly can make the required behaviour clearer. But the instruction still has to be followed, and the system still needs to recognise where the gaps are.

We should therefore judge an admission of uncertainty by its substance. Does it identify the unresolved point? Does it explain what evidence would settle it? Has the assistant separated established information from inference?

A useful response might be: the supplied document establishes when the payment was sent, but does not identify when it was received. That tells the reader exactly where the evidence stops.

Compare that with a confident date followed by a general warning that AI can make mistakes. The warning transfers work to the reader without identifying the weakness in the answer.

The next step

The failure to say "I don't know" has several possible ingredients: imperfect uncertainty signals, incentives that favour attempting an answer, feedback that rewards persuasive agreement, and questions that carry misleading assumptions.

Human reluctance to acknowledge an error can then magnify the consequences. The system need not feel embarrassment, or consciously decide to conceal anything, for its output to assist somebody who is resisting a correction.

Developers have worked on these problems. Some interventions improve what a model knows. Others change when it answers, how it checks or how it communicates its limitations.

Part 3 examines those developments: **What Has Changed Since ChatGPT First Appeared?**

Sources and further reading

1. *Mata v Avianca, Inc.*, 678 F Supp 3d 443 (SDNY, 22 June 2023), particularly Findings of Fact [45] – [47] and Conclusions of Law [21] – [24]. Judgment.

2. Kadavath et al. (Anthropic), *Language Models (Mostly) Know What They Know* (2022), particularly §§4–5 on self-evaluation and prediction of answerability, and §6.1 on limitations. Paper.

3. Orgad et al., *LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations*, ICLR 2025 (first circulated 2024). The findings concern trained probes, specified models and datasets, rather than a general-purpose truth detector. Paper.

4. Kalai et al. (OpenAI and Georgia Tech), *Why Language Models Hallucinate* (2025), particularly §1.2 and §4.1 on evaluation incentives and binary scoring. This provides an organising explanation, not an established cause of every individual error. Paper.

5. Ouyang et al. (OpenAI), *Training Language Models to Follow Instructions with Human Feedback* (2022), particularly §3.1 on methodology and §4 on results. Paper.

6. Wei et al. (OpenAI), *Measuring Short-Form Factuality in Large Language Models* (2024), particularly §2.4 on grading and metrics and §4 on stated confidence. Its deliberately challenging questions do not establish ordinary-use error rates. Paper.

7. Sharma et al. (Anthropic-led), *Towards Understanding Sycophancy in Language Models*, ICLR 2024 (first circulated 2023), particularly §3.2 on responses to challenges and §4 on human preferences. Paper.

8. Ibrahim, Hafner and Rocher, *Training Language Models to Be Warm Can Reduce Accuracy and Increase Sycophancy*, *Nature* 652, 1159–1165 (2026), including the discussion of limitations. Full text.

9. OpenAI, *Sycophancy in GPT-4o* (29 April 2025) and *Expanding on What We Missed with Sycophancy* (2 May 2025). These are the developer's accounts of the incident and its contributing factors. Initial account · Further explanation.

10. *In re Sanctions Order of Kenney, Kerry*, No 25-C-389 (La App 5 Cir, 23 October 2025), printed pp 2–5 and 18–20, particularly "AI-generated Case Summaries". Judgment.

11. *In re Domestic Partnership of Campos & Munoz*, No D085584 (Cal Ct App, 5 March 2026, subsequently modified), printed pp 5 – 12. Judgment with modification.