Why AI Makes Things Up

Part 1 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

In *Mata v Avianca*, ChatGPT supplied fictitious judgments and then assured a lawyer that the cases could be found on Westlaw and LexisNexis. The extra detail made the answer sound easier to verify. It still came from the system that had generated the cases.[1]

A language model learns to predict what comes next in text. That helps it reproduce the form of a judgment, a research citation or a report. A plausible continuation can still attach a real paper to the wrong author or add a cause that the source never gave.[3] [4] [6]

There is more than one route to error. The model may have little basis for answering, repeat a misconception, combine genuine details incorrectly or mishandle evidence already in front of it. An initial mistake can also shape the explanation that follows, making the answer more coherent as it becomes less accurate.[3] [4] [6] [7] [8]

When an answer matters, check the claim against its source. The system's reassurance may sound persuasive, especially when some details are right. The evidence has to come from somewhere you can inspect.

In 2023, lawyers pursuing a claim against the airline Avianca submitted legal authorities generated by ChatGPT to a federal court in New York. The material looked like legal research: case names, citations, quotations and explanations of how the decisions supported their argument.

Six of the cited decisions did not exist.[1]

When the court required copies, purported judgments were supplied. These too contained inventions. One fictitious decision attributed an opinion to real judges, borrowed identifying details from other proceedings and cited further authorities that either did not exist or failed to support the propositions attributed to them.[1]

The court also records that, when asked whether its cases were real, ChatGPT assured the lawyer that they were and could be found on Westlaw and LexisNexis. The judgment identifies conflicting accounts of when he sought that reassurance.[1]

Courts in England and Wales have faced related failures. In 2025, the Divisional Court used the joined *Ayinde* and *Al-Haroun* cases to warn the profession about unchecked AI-assisted research.[2] Part 2 returns to those cases.

The judgment in *Mata v Avianca* records serious failures by the lawyers responsible. But it also presents a puzzle that extends well beyond legal practice.

How can a system reproduce the outward form of knowledge so convincingly while inventing the substance?

The familiar explanation is that AI "predicts the next word". That is a starting point. To understand why hallucinations happen, we need to follow the process a little further.

What are we calling a hallucination?

The word is an awkward borrowing. In human medicine, a hallucination concerns perception. A language model producing an imaginary court judgment is doing something different.

In AI research, the term also covers several kinds of failure. A model may invent a publication, attach the wrong date to a real event, distort a document it has been asked to summarise, or add information that the document never contained.[3] [4]

Those distinctions matter.

Suppose a report says that a company's revenue increased. A summary adds that the increase resulted from a successful advertising campaign. Perhaps that explanation is true. But if the report gives no explanation, the summary has introduced an unsupported claim. It has crossed the boundary between reporting the document and supplying its own account.

Researchers were investigating such failures in summarisation before ChatGPT appeared. The problem was already clear: fluent text could depart from the evidence it was meant to represent.[4]

For this series, the central concern is generated material presented as factual or source-based when it is false, invented or unsupported. A request to write a fictional detective story is a different task. We asked for the invention.

How text becomes a model

During a major stage of training, a language model encounters large quantities of material and learns to predict the next *token*. A token can be a word, part of a word or punctuation.

Its predictions are compared with the text that actually follows. Training adjusts numerical parameters so that subsequent predictions improve.[5]

Consider an invented training example:

The committee postponed its decision until…

Predicting what comes next requires sensitivity to grammar, context and the kinds of things committees do. Across sufficiently varied material, successful prediction also benefits from learning relationships between people, places, events and ideas.

That is why describing a capable model as "just autocomplete" leaves so much unexplained. The training objective is relatively simple to state; the abilities acquired through pursuing it can be complex.[3] [5]

Further training then shapes the model into an assistant: following instructions, answering questions and displaying behaviours its developers seek to encourage.[5]

There is, however, a crucial distinction. Learning to produce an appropriate continuation does not establish that every factual statement in that continuation has been checked.[3]

Generating an answer is different from retrieving a record

Imagine asking for the title of a paper written by a particular scientist.

A conventional database might retrieve a stored author–publication record. A language model answering from its learnt parameters generates a response using the patterns and relationships represented in those parameters.

It may produce the correct title. It may reproduce information encountered during training. But the process can also combine genuine elements incorrectly: a real scientist, a real research field and a real paper, joined by an authorship claim that is false.[6]

A separate search or retrieval system can supply records for the model to consult. We will examine that later in the series. For now, the important point is that an answer generated from learnt knowledge does not automatically come with an inspected source behind each statement.

Think of a hypothetical person who remembers the subject of a book, its publisher's style and roughly when it appeared, but cannot recall its title. They might construct a remarkably convincing title from those fragments.

The comparison illustrates the distinction between familiarity and accurate recall. It does not imply that a model experiences human remembering.

There is more than one route to a false answer

Sometimes the required information is missing or too weakly learnt to be recovered reliably. An obscure date or an arbitrary identifying detail may offer little basis for inference.[3]

Sometimes the model has learnt a falsehood. Human writing contains misconceptions, superstition and confidently repeated mistakes. Research using the TruthfulQA benchmark examined models' tendency to reproduce familiar human falsehoods. An incorrect answer need not have been invented afresh.[7]

Sometimes genuine knowledge is combined incorrectly. Knowing that a researcher works in deep learning does not establish which deep-learning papers they wrote.[6]

And sometimes the relevant evidence is available, but the model mishandles it. A summary can attribute a statement to the wrong person or turn a qualified observation into an unconditional conclusion.[4]

These routes should not be collapsed into one explanation. "The training data was wrong" cannot explain every invented detail. Equally, an error does not prove that the model lacked all relevant information.

The failure may concern how available information was selected, connected or used.[3] [8]

A glimpse inside the machinery

One particularly revealing study by Anthropic researchers went beyond examining answers and investigated internal mechanisms in its Claude 3.5 Haiku model.

Asked to name a paper by the AI researcher Andrej Karpathy, the model attributed *ImageNet Classification with Deep Convolutional Neural Networks* to him. The paper exists, but Karpathy was not one of its authors.

The researchers identified activity associated with recognising familiar entities and a circuit that appeared to promote declining to answer by default. Their analysis suggested that familiarity with Karpathy's name could suppress that default even though the required authorship relationship was missing. Intervening on the identified features could change the model's behaviour, eventually producing a refusal.[6]

In accessible terms, recognising the person helped open the door to answering; it did not ensure that the specific answer was known.

This is a bounded finding from selected examples in one model, using an incomplete account of its internal computations. It does not establish a universal switch behind every hallucination.[6]

It nevertheless provides something more informative than "AI sometimes gets confused": an investigated mechanism through which familiarity can permit an unsupported answer.

The first mistake can recruit further mistakes

Once a model begins answering, its own output becomes part of the context for what follows.

That creates an opportunity for an initial error to grow.

A hypothetical answer might wrongly state that a particular researcher won an award. The next sentences may then supply a year, describe the supposed award-winning contribution and explain its significance. Each addition makes the account more substantial without repairing the original mistake.

Research on "hallucination snowballing" tested related behaviour in GPT-3.5 and GPT-4. On three constructed question-answering datasets, the models produced incorrect answers with false supporting explanations. When questioned separately, GPT-3.5 identified about 67% of the false supporting claims as incorrect; GPT-4 identified about 87%. These are recognition rates for the false claims examined, not general accuracy rates.[8]

That does not establish that every explanation is an invention. It shows why an explanation attached to an answer cannot automatically serve as an independent check on it.

The exchange in *Mata* illustrates the practical danger: asking the system that supplied a doubtful authority to vouch for it is not independent verification.[1]

An account can become more coherent as it becomes less accurate.

Is randomness responsible?

Randomness can influence generation. Depending on the settings, a model may sample from possible next tokens rather than always select the highest-ranked one.

But eliminating sampling does not eliminate factual error.

The main snowballing experiments used *greedy decoding*: selecting the most probable next token at each step. The models still generated false answers and false explanations. The published paper also reports that higher-temperature sampling and beam search did not remove the phenomenon.[8]

A mistaken association can be the model's preferred continuation. Making that choice more repeatable does not make the association correct.

GEEK BOX: Likely language and factual truth

A model assigns probabilities to possible continuations of text. Those probabilities depend on the context and what it has learnt.

The probability assigned to a token is not, by itself, a measured probability that the resulting factual claim is true.[3] [5]

Consider an invented request for the date on which a little-known organisation was founded. A particular year may fit the model's learnt associations and the developing sentence. Establishing that year as correct requires a further basis: reliable knowledge, suitable evidence or a successful verification process.

This distinction does not mean prediction is unrelated to truth. Learning from accurate material can support accurate answers. It means that selecting plausible text and certifying a proposition are different operations.

Why the result is so persuasive

The fictional legal authorities in *Mata* illustrate a dangerous combination: recognisable form and unreliable content.

A reader accustomed to legal materials sees names, reporter references, quotations and judicial language. Those are familiar signals of a document that can be located and checked. Here, the signals survived while parts of the underlying authority disappeared.[1]

The same difficulty can arise with a scientific reference, a historical anecdote or a polished summary. The reader encounters the result in a form normally associated with researched work.

This is my practical interpretation of the evidence: the risk lies partly in our habit of treating competent presentation as evidence of competent verification. With generated text, that connection needs to be tested.

The problem is especially difficult when an answer mixes correct and incorrect material. Recognising several true details may reassure us about the next detail, precisely where independent checking becomes necessary.

What this explanation establishes

Next-token prediction helps explain how convincing unsupported text can be produced. The fuller account includes uneven knowledge, learnt misconceptions, incorrect associations, failures to follow evidence and the influence of earlier generated text.[3] [4] [6] [7] [8]

It also leaves a question.

If the system lacks a reliable answer, why does it continue? Why not identify the gap and stop?

The answer involves uncertainty, training and the incentives created by evaluation. Models can sometimes contain useful signals of whether an answer is likely to be correct without consistently acting on those signals or communicating them faithfully.[3] [6] [8]

That is the subject of Part 2: **Why Doesn't It Just Say "I Don't Know"?**

Sources and further reading

1. *Mata v Avianca, Inc.*, 678 F Supp 3d 443 (SDNY, 22 June 2023), particularly Findings of Fact [25] – [29] on the fabricated *Varghese* decision, and [45] – [47] on ChatGPT's assurances and the conflicting accounts of their timing. Judgment.

2. *R (Ayinde) v London Borough of Haringey; Al-Haroun v Qatar National Bank* [2025] EWHC 1383 (Admin), 6 June 2025, particularly [3] and [7] – [9]. Judgment.

3. Kalai et al. (OpenAI and Georgia Tech), *Why Language Models Hallucinate* (2025), particularly §§1 and 3. The explanation above is a synthesis, rather than a claim that the paper establishes every causal route in every deployed system. Paper.

4. Maynez et al., *On Faithfulness and Factuality in Abstractive Summarization* (2020), particularly §2 on the distinction between faithfulness and factuality, and the analysis of generated summaries. Paper and PDF.

5. OpenAI, *GPT-4 Technical Report* (2023), introduction and §2; Brown et al., *Language Models are Few-Shot Learners* (2020), particularly §2. GPT-4 report · Brown et al..

6. Lindsey et al. (Anthropic), *On the Biology of a Large Language Model* (27 March 2025), "Entity Recognition and Hallucinations", especially "Case Study of a Natural Hallucination: Academic Papers", and the limitations section. Research.

7. Lin, Hilton and Evans, *TruthfulQA: Measuring How Models Mimic Human Falsehoods* (2021; ACL 2022), particularly the introduction and §4.3. Its historical experiments do not establish present-day product error rates. Paper.

8. Zhang et al., *How Language Model Hallucinations Can Snowball*, ICML 2024, PMLR 235, pp 59670–59684 (preprint first released in 2023), particularly §§2–4. These are findings about constructed tasks and historical model versions. Published paper.