AI LITERACY SERIES

AI hallucinations: why does it make things up?

A plain-English guide to one of the most misunderstood properties of AI

Ryan Bishop | May 2026 | ryanbishop.co.uk

1. Why AI makes things up

If you have spent any time using a large language model, you have probably encountered a response that seemed authoritative, was well-written, and turned out to be wrong. Perhaps a date was incorrect. A citation did not exist. A product detail was invented. A person's history was fabricated with apparent confidence.

This is hallucination. And it is not a glitch, a bug, or evidence that the model is broken. It is a predictable, documented, structural property of how these systems work.

Episode 4 explained that large language models operate through pattern matching at extraordinary scale — producing responses that reflect what they learned to produce, rather than reasoning through problems or retrieving verified facts. Hallucination is what happens when that pattern-matching process leads somewhere that sounds right but is not.

Understanding this is one of the most practically important things anyone using these tools can know. Not because it should stop you using them — but because knowing the mechanism is what allows you to use them responsibly.

2. What hallucination actually means

The term has a precise meaning in the research literature. Ji et al. define hallucination in natural language generation as the production of text that is **unfaithful or nonsensical relative to the source** — content that gives the impression of being grounded in real context, although it is actually hard to specify or verify.

The parallel to psychological hallucination is deliberate. In psychology, a hallucination is a perception that feels real in the absence of an appropriate stimulus. In AI, hallucinated text feels authoritative in the absence of factual grounding. Both are hard to tell apart from the real thing at first glance.

Researchers distinguish two types, and the distinction matters in practice.

Intrinsic hallucination

The model produces output that directly contradicts what it was given or what is known to be true. A summary that states an approval happened in 2021 when the source clearly states 2019. A name that is wrong. A number that conflicts with the data provided.

Extrinsic hallucination

The model produces output that goes beyond what can be verified from the source — information that is neither supported nor contradicted by what it was given. This is the more common and more insidious type. The model adds a detail, a reference, a fact that simply does not exist in the source — and there is no obvious signal that anything has gone wrong.

Extrinsic hallucination is not always incorrect. Occasionally the additional information happens to be factually accurate, drawn from the model's training. But because it cannot be verified against the source, it carries real risk — particularly in professional contexts where precision matters.

The definition in plain English

Intrinsic hallucination: the model contradicts what it was told or what is true.

Extrinsic hallucination: the model adds information that cannot be verified from the source.

Both types share the same characteristic: the output reads as fluent and authoritative regardless of whether it is accurate.

3. Why it happens — and why it cannot simply be fixed

Hallucination is not caused by one thing. Ji et al. identify three distinct contributors, and together they explain why hallucination is a structural property of these systems rather than a correctable error.

What the model learned from

Large language models are trained on enormous quantities of text. That text is not perfectly consistent or perfectly factual — it reflects the diversity, noise, and occasional inaccuracy of the sources it came from. When training data contains inconsistencies between source and reference, the model learns patterns that do not always stay faithful to the source. The divergence is baked in before the model is ever deployed.

There is a further complication. Pre-trained models memorise knowledge in their parameters — statistical representations of what they repeatedly encountered during training. When asked to produce output, a model draws on this parametric knowledge alongside the information it has been given. Ji et al. document that models tend to prioritise their parametric knowledge over the specific input provided. The result: the model's learned assumptions can override the facts in front of it.

How the model generates

The generation process itself contributes to hallucination. At each step of producing a response, the model selects what comes next based on statistical likelihood — what tends to follow, given everything before it. This process has no built-in truth-check. It does not verify the output against a source of fact. It completes the pattern.

The decoding strategies that make models more useful — those that introduce some variation to avoid repetitive or bland responses — also increase hallucination. More diversity in generation means a higher probability of producing content that sounds plausible but diverges from the truth. The trade-off between usefulness and faithfulness is real, and it is not trivially resolved.

The absence of a truth-checking mechanism

A person who does not know the answer to a question can recognise that uncertainty and say so. A large language model completing a pattern has no equivalent check. It produces the most statistically plausible continuation — whether or not that continuation is accurate. The model has no mechanism for knowing when it does not know.

This is not a design oversight that will eventually be patched. It is a consequence of how these systems learn and generate. Mitigation methods exist — retrieval augmentation, fine-tuning on faithful datasets, post-processing checks — and they reduce hallucination meaningfully. But they do not eliminate it. Hallucination is, as Ji et al. put it, an artefact of natural language generation.

Why it cannot simply be fixed

Hallucination arises from three sources simultaneously: the characteristics of training data, the model's tendency to prioritise learned knowledge over provided input, and the statistical nature of the generation process itself.

Each can be partially addressed. None can be fully eliminated. Understanding this is what separates informed use of AI from naive reliance on it.

4. Why it is hard to catch

Hallucinated content does not look wrong. That is the problem.

It is produced by the same process as accurate content. It uses the same sentence structures, the same register, the same apparent confidence. It does not flag itself as uncertain. It does not come with a caveat. A well-written hallucination is indistinguishable from a well-written accurate response — unless you verify it.

This matters most where the output is long or complex. When a model produces a detailed analysis, a structured summary, or a multi-part answer, errors are distributed through a document that otherwise reads as coherent and authoritative. The reader's attention is on the argument, not on verifying each component fact.

The practical implication is straightforward: fluency is not evidence of accuracy. A confident, well-written response from a large language model is not a substitute for verification. This is not a counsel of despair — it is the same standard we apply to any information source. It just needs to be applied consciously, because the fluency of these systems makes it easy to forget.

5. For families — teaching children to fact-check AI

The concept of hallucination is, in some ways, easier to explain to children than to adults. Children have fewer ingrained assumptions about what computers can and cannot do. They are often more willing to accept that a tool can be impressive and unreliable at the same time.

The key insight to build is this: **AI is not looking things up. It is producing what it learned to produce.** And what it learned to produce is not always accurate.

Start with what they already know

Most children have encountered autocomplete — on a phone, in a search bar, in a word processor. Autocomplete suggests what usually comes next, based on what it has seen before. It does not know what you actually mean to say. Large language models work on a similar principle, at vastly greater scale and sophistication. They predict what a good response looks like. They do not verify that it is true.

The homework help conversation

If your child uses AI tools for homework — for research, for checking facts, for summarising topics — the habit to build is the same habit a good teacher would want: check it. Not because AI is usually wrong, but because it is sometimes wrong in ways that are hard to spot.

Ask them: if a friend told you a fact with total confidence, would you put it in your homework without checking? The same question applies here. Confidence is not the same as accuracy, whether it comes from a person or a model.

Critical AI use in practice

Three questions worth teaching children to ask about any AI-generated content:

- Does this match what I already know? If the answer contradicts something established, that is a signal to verify.
- Can I find a source for this? If it is a specific claim — a date, a statistic, a quotation — it should be traceable to a primary source.
- Is this the kind of question where precision matters? For creative or exploratory tasks, accuracy is less critical. For facts that will be cited or acted on, it is essential.

A conversation starter for families

Ask your child: if someone wrote an essay with complete confidence and perfect grammar, does that mean everything in it is true?

That is the most important thing to understand about AI-generated text. The writing quality tells you nothing about the factual accuracy. They are independent properties — and AI systems have learned to produce the first very well, without any guarantee of the second.

6. For organisations — managing hallucination risk in practice

In professional contexts, hallucination is not an abstract concern. It is a risk management question. The stakes vary by use case — and calibrating your controls to the stakes is the starting point for responsible deployment.

High-stakes outputs require human verification

Any AI-generated content that will be submitted, published, cited, or acted on in a high-stakes context needs human review at the level of individual claims, not just overall coherence. This applies to regulatory submissions, clinical summaries, payer dossiers, legal documents, and financial reports.

The risk is not that AI produces obviously wrong content. It is that it produces subtly wrong content — a date adjusted, a citation invented, a statistic misremembered — embedded in a document that otherwise reads correctly. Reviewing for overall quality is not sufficient. Claim-level verification is required.

The citation problem

Hallucinated citations are among the most professionally damaging failure modes. A model asked to support an argument with references will sometimes produce citations that do not exist — author names, journal titles, and DOIs that are plausible in structure but entirely fabricated. They look real. They are formatted correctly. They do not exist.

In pharmaceutical communications, this is not a theoretical risk. A hallucinated citation in a submission or a medical education piece is a compliance failure. Every reference produced by an AI tool in a professional document should be verified against the primary source before use.

Retrieval augmentation reduces — but does not eliminate — the risk

Many enterprise AI deployments now use retrieval-augmented generation: connecting the model to a specific document library or database, so that it draws on verified sources rather than solely on its training. This meaningfully reduces hallucination by anchoring the model's responses to specific, controllable content.

It does not eliminate it. The model still processes and presents retrieved information through the same pattern-matching mechanism. It can misrepresent, misinterpret, or inappropriately combine what it found. Retrieval reduces the problem; it does not resolve it. Human oversight remains necessary.

Hallucination risk scales with task complexity

Shorter, factual tasks — extracting specific data from a document provided in full, reformatting structured content, summarising a clearly bounded source — carry lower hallucination risk. The model has a specific, verifiable source to work from.

Longer, generative tasks — synthesising across multiple sources, generating arguments, producing content from memory rather than provided input — carry higher risk. The model is drawing more heavily on its parametric knowledge and has more opportunity to introduce unverifiable content.

Calibrate your verification effort to the task type, not just the stakes of the output.

The practical principle for organisations

Hallucination is a known, structural property of large language models. It cannot be assumed away, and it cannot be entirely engineered out.

The appropriate response is not to avoid AI — it is to build verification into workflows wherever the accuracy of specific claims matters. The question is not whether the AI might be wrong. It is whether your process will catch it when it is.

Sources & Reference

This episode draws on peer-reviewed literature in natural language processing and AI, combined with original analysis by Ryan Bishop.

Primary source — hallucination taxonomy and contributors: Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A. and Fung, P. (2023). 'Survey of Hallucination in Natural Language Generation.' *ACM Computing Surveys*, 55(12), Article 248. https://doi.org/10.1145/3571730

Definition of hallucination used in this episode: Ji et al. define hallucination in natural language generation as content that is "nonsensical or unfaithful to the provided source content" — text that gives the impression of being grounded in real context, although it is actually hard to specify or verify (p. 248:3).

Intrinsic and extrinsic hallucination: The two-category taxonomy follows Ji et al. (2023), Section 2.1. Intrinsic hallucination denotes output that contradicts the source; extrinsic hallucination denotes output that cannot be verified from the source (p. 248:3).

Contributors to hallucination: The three contributors discussed in this episode — source-reference divergence in training data, parametric knowledge bias, and erroneous decoding — are drawn from Ji et al. (2023), Section 3.

Hallucination as an artefact of NLG: Ji et al. (2023), Section 13: "Hallucination is an artifact of NLG and is of concern because they appear fluent and can therefore mislead users" (p. 248:31).

Want to discuss what responsible AI implementation looks like for your organisation? Get in touch: ryan@ryanbishop.co.uk

RAI Disclosure: This content was produced with AI assistance, reviewed and contextualised by Ryan Bishop. The process is intentional — this is augmentation in action.