CUADERNO DE IDEAS

Why Does Artificial Intelligence Make Things Up? H

Tecnología e IA · 09/11/2026 · Por Camilo Calderón

Artificial intelligence can produce convincing answers that contain false facts, nonexistent references, or incorrect explanations. This is known as an AI hallucination. Discover why it happens and how to verify information.

Why Does Artificial Intelligence Make Things Up? How AI Hallucinations Work

Artificial intelligence can explain complex concepts, summarize documents, analyze information, and answer questions in seconds. However, it can also invent facts, attribute quotes to people who never said them, or present scientific references that simply do not exist.

The most curious part is that it can do this with flawless writing and a completely confident tone.

This phenomenon is known as an artificial intelligence hallucination and represents one of the most important challenges for the reliability of language models.

But why does it happen? How can technology capable of processing enormous amounts of information produce false answers? And, most importantly, what can we do to avoid being misled?

To answer these questions, we need to understand how generative AI works and what its limitations are.

What Are Artificial Intelligence Hallucinations?

In artificial intelligence, a hallucination occurs when a model generates false or inaccurate information, or information that is not supported by available sources, but presents it as a valid answer.

Imagine asking a chatbot for three scientific studies on the effects of a particular technology on the brain.

The system might give you the researchers names, article titles, the journals where they were supposedly published, and even their DOI identifiers.

At first glance, everything looks correct. But when you try to find the documents, you discover that some of them never existed.

Something harder to detect can also happen: the study does exist, but the AI misinterprets its findings or attributes conclusions to it that the authors never made.

Hallucinations can take different forms:

Invented facts: incorrect dates, statistics, names, or events.

Nonexistent references: fabricated books, scientific articles, authors, or links.

Distorted information: real facts presented out of context.

Unsupported conclusions: claims that sound reasonable but lack sufficient evidence.

The fundamental problem is that a well-written answer is not necessarily a true one.

How Does a Language Model Work?

To understand why hallucinations happen, we first need to know how large language models, or LLMs, work.

These systems are trained on large amounts of text to learn patterns, relationships between words, grammatical structures, and associations among concepts.

When generating an answer, the model calculates which language elements are most likely to continue the context and builds the text progressively.

For example, if we write:

â€œWater boils at a temperature of...â€?

A model trained on enough information will probably continue with â€œ100 degrees Celsius,â€? because it has learned that this sequence frequently appears in contexts about the boiling point of water at standard atmospheric pressure.

However, generating a linguistically likely continuation is not the same as consulting a source and checking every claim.

This distinction is essential: language models are designed to generate coherent answers, but that does not mean they automatically verify the truth of everything they produce.

Why Does AI Invent Information?

There is no single cause. Hallucinations can result from different limitations in the model, its data, and the processes used to train and evaluate it.

1. Predicting Words Is Not the Same as Checking Facts

A model can learn how scientific explanations are usually written without independently verifying every fact it includes.

That is why, when asked for a bibliographic reference, it can generate a convincing combination of names, titles, and dates that follows the usual structure of an academic publication, even if the reference is false.

The answer may look professional because it reproduces patterns learned during training.

But the appearance of rigor does not replace evidence.

2. Incomplete Information or Information Outside Its Knowledge

Models have limitations related to the data used during training and the information available when they respond.

If you ask about a recent event, a highly specific study, or an obscure fact, the system may not have enough information.

In some circumstances, instead of recognizing this limitation, it may generate an answer based on learned associations.

This explains why questions about recent events, specialized research, and very specific facts require special caution.

3. The Tendency to Guess Under Uncertainty

A study published by OpenAI in September 2025, titled Why Language Models Hallucinate, examines a particularly interesting explanation.

The authors argue that certain training and evaluation procedures may encourage models to guess an answer instead of acknowledging that they do not know something.

We can understand this through a comparison.

Imagine an exam where leaving a question blank always earns zero points, while guessing gives you some chance of getting it right.

The student would have an incentive to answer even when they do not know the solution.

Something similar can happen when models are evaluated using criteria that reward correct guesses but do not sufficiently value acknowledging uncertainty.

This does not mean that every model always responds this way, but it helps explain one mechanism that contributes to the problem.

4. Ambiguous Questions and Lack of Context

The way we phrase a question can also affect answer quality.

For example, asking â€œWhat does science say about artificial intelligence?â€? is too broad.

We have not specified which AI application interests us, which population we want to study, or what outcomes we are looking for.

The model has to interpret our intention and construct an answer from limited context.

Providing more precise instructions, narrowing the topic, and sharing reference documents can help reduce incorrect interpretations.

However, even a well-written question does not guarantee a completely correct answer.

Why Do False Answers Seem So Convincing?

Here we find one of the most delicate aspects of hallucinations.

Language models can produce explanations with excellent grammar, specialized vocabulary, well-organized paragraphs, and a professional tone.

These qualities can create an impression of authority.

However, we need to distinguish between two concepts:

Fluency: how natural and well-structured an answer seems.

Factuality: how accurately it represents the facts and available evidence.

An answer can be very fluent and still contain important errors.

This affects how we use technology. When an explanation sounds convincing, we may accept its content without checking it.

The risk increases when we use these tools to prepare academic work, make financial decisions, interpret legal information, or seek health guidance.

The confidence conveyed by an answer should not be confused with its truthfulness.

How Can We Detect an AI Hallucination?

There is no visual or linguistic sign that can identify every invented answer. A false claim can be written just as well as a true one.

That is why the best strategy is to build verification habits.

1. Check the Original Sources

If the AI mentions a scientific study, look for its title, authors, and original publication.

It is not enough for the system to provide a reference that looks complete.

Confirm that the document exists and that it actually supports the claim you want to use.

2. Cross-Check Important Facts

When an answer includes numbers, dates, statistics, or important claims, compare them with reliable independent sources.

For scientific topics, prioritize academic papers, specialized journals, and research organizations.

3. Ask It to Separate Facts From Interpretations

You can ask the model to separate facts supported by sources from interpretations or hypotheses it is developing.

This makes it easier to review the reasoning and identify claims that need more checking.

4. Use Search and Retrieval Tools

Some systems can consult documents or external sources before generating an answer.

A technique known as RAG (Retrieval-Augmented Generation) provides information retrieved from selected documents to support the answer.

This can improve accuracy and reduce certain errors, but it does not guarantee that hallucinations will disappear.

The model may still misinterpret a source, use outdated information, or draw conclusions that do not follow from the documents it consulted.

5. Learn to Question Overly Confident Answers

An answer stated firmly is not necessarily true.

If the information matters, ask for verifiable evidence. If the system cannot support a claim,

it is reasonable to remain uncertain instead of accepting it as fact.

Can Artificial Intelligence Hallucinations Be Eliminated?

Reducing them is one of the major challenges in artificial intelligence research.

Different strategies are currently being studied to improve model reliability:

Improving the quality of training data.

Developing ways to estimate uncertainty.

Adding external sources and verifiable documents.

Evaluating the faithfulness of answers.

Improving the ability to recognize when information is insufficient.

A study published in Nature in June 2024, titled Detecting Hallucinations in Large Language Models Using Semantic Entropy, examined statistical methods for estimating answer uncertainty.

The researchers analyzed how differences in meaning among several answers generated by the same model can help identify situations with a higher likelihood of incorrect information.

This research matters because it is not enough to make AI generate more elaborate answers. We also need systems that can recognize their own limitations.

The goal should not be only to make artificial intelligence answer more, but to help it better distinguish what it can support, what it is inferring, and what it does not have enough information to answer.

Artificial Intelligence Is a Tool, Not an Infallible Source

Hallucinations do not mean artificial intelligence is useless. They mean we should understand its capabilities and limitations before trusting its results.

It can be an extraordinary tool for research, learning, organizing ideas, and exploring problems. But we should not turn its answers into automatic truths.

Our responsibility to verify information becomes especially important when we use these technologies to make decisions or communicate knowledge to other people.

The next time an AI gives you an apparently perfect answer, remember this:

An explanation sounding convincing does not mean reality supports it.

Learning to use artificial intelligence also means learning to question it.

Further Reading

Wang, Y. et al. (2024). Factuality of Large Language Models: A Survey. Proceedings of EMNLP 2024. https://doi.org/10.18653/v1/2024.emnlp-main.1088

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting Hallucinations in Large Language Models Using Semantic Entropy. Nature, 630, 625â€“630. https://doi.org/10.1038/s41586-024-07421-0

Kalai, A. T. et al. (2025). Why Language Models Hallucinate. OpenAI Research.

https://openai.com/index/why-language-models-hallucinate/

Soy Camilo CalderÃ³n

Learn. Discover. Evolve.