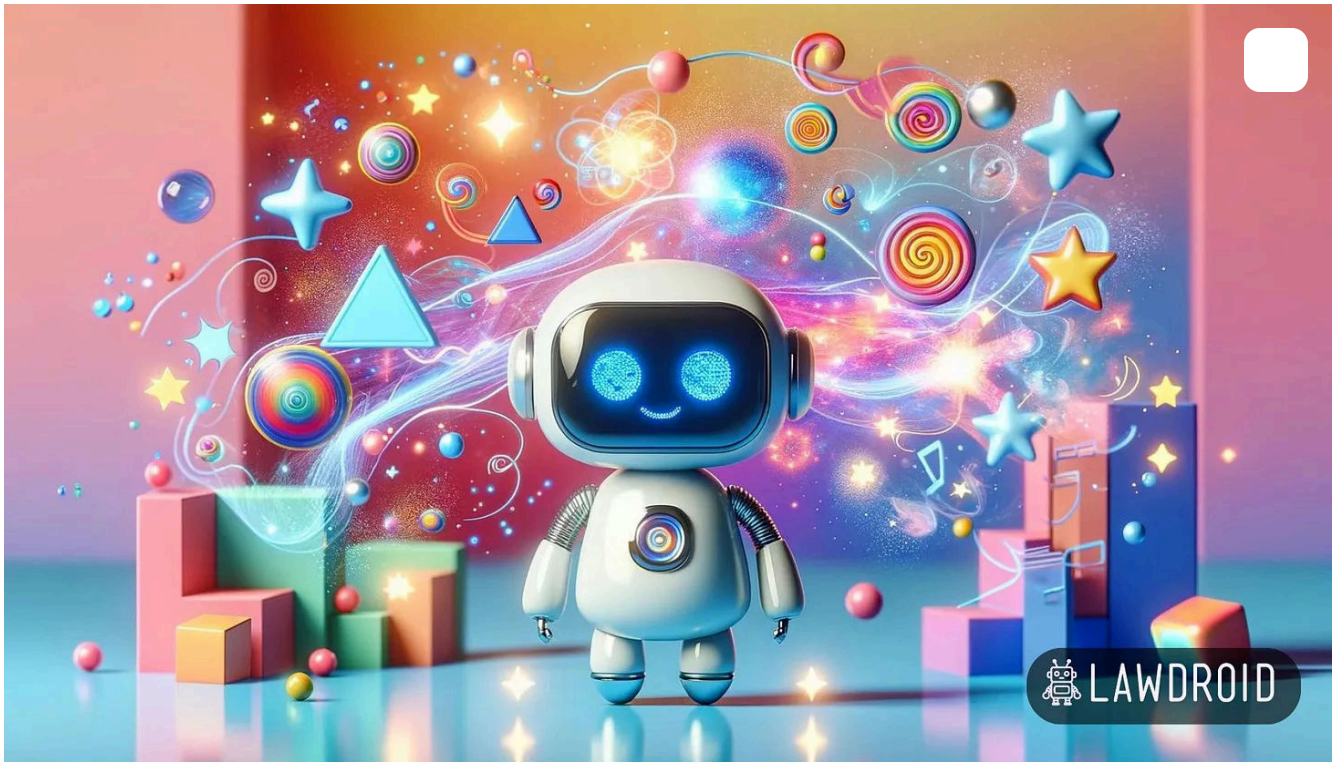
Hallucinations: What Are They, Why Do They Happen, How to Fix Them?

Where I explain what hallucinations are in the context of large language models

TOM MARTIN

MAY 08, 2024



Thanks again to all my readers! Appreciate you. ❤️

And to you newbies, as John Oliver would say: welcome, welcome, welcome! 👋

This substack, LawDroid Manifesto, is here to keep you in the loop about the intersection of AI and the law. Please share this article with your friends and colleagues and remember to tell me what you think in the comments below.

If you've heard the term "hallucination" used in reference to large language models (LLMs), you've probably asked yourself what that means, why it happens, and if there's any way to fix it. Well, you're in luck: this article is a practical guide to AI hallucinations. Too many explanations just scratch the surface and I want to sort through the nuts and bolts. Or, maybe I'm hallucinating? How about that?

If this sounds interesting to you (then you have qualified as nerd adjacent for sure), please read on...

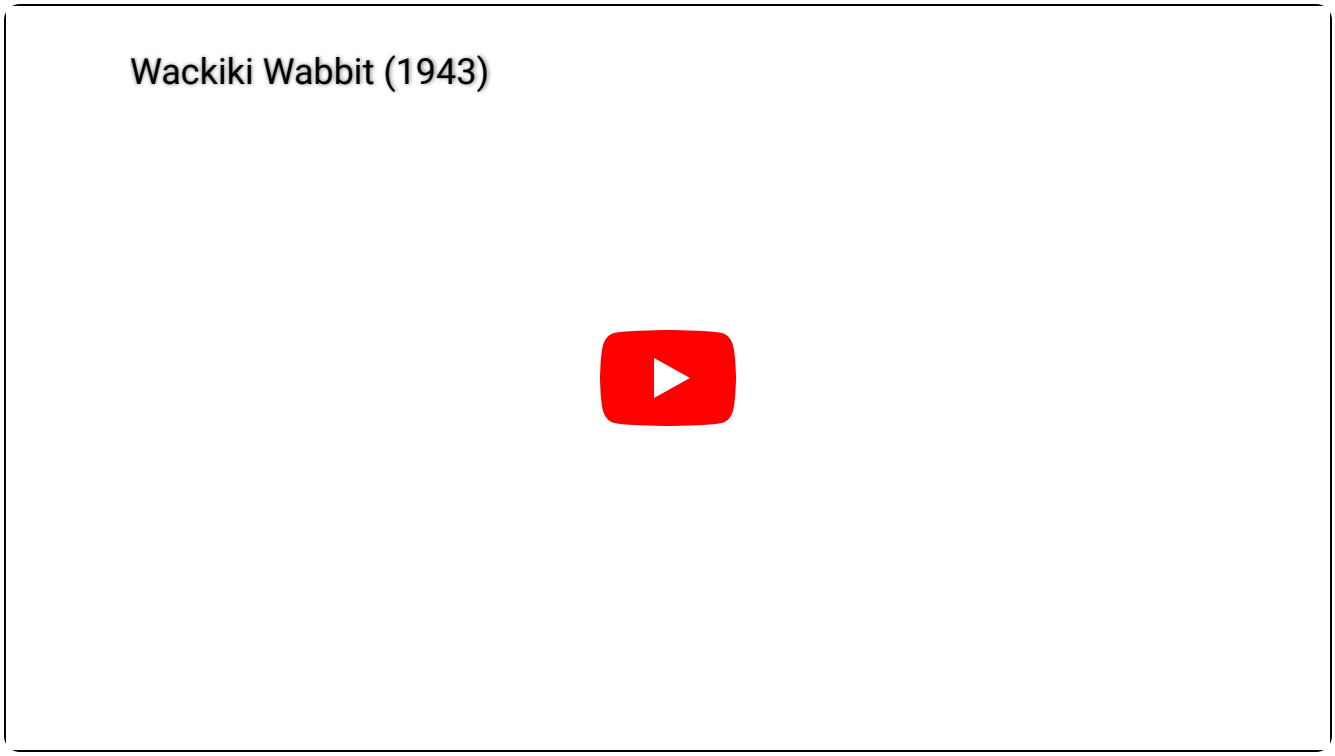
What Are Hallucinations?

As lawyers we love definitions so let's start with one:

Hallucination

> the experience of seeing, hearing, feeling, or smelling something that does not exist, usually because of a health condition or because you have taken a drug. [1]

When I think of hallucinations, I picture Bugs Bunny stuck on a deserted island with two hungry castaways, who mistake each other for a tasty hot dog or hamburger, as in the Looney Tunes episode, "Wackiki Wabbit." Never heard of it? Well, you my friend are missing out! For your viewing pleasure, here you go:



Be it hunger, dehydration, delirium, or drugs, when hallucinations happen to human beings, you definitely know that something has gone wrong. It is not normal for people to hallucinate.

Yet, when large language models “hallucinate,” it is (as I will explain later) a result of the LLM’s normal operation. To this extent, this anthropomorphism of artificial intelligence gets it backwards. It also unnecessarily stigmatizes the phenomenon with the pall of mental illness or drug abuse.

The definition of hallucinations for LLMs generally goes something like this:

AI Hallucination

> AI hallucinations are incorrect or misleading results that AI models generate. [2]

Now, I said the definition “goes something like” the above, because to date there is no precise and universally accepted definition. [3] For example, in the context of text translation, AI hallucinations have been described alternatively as "fluent but irrelevant,” “abnormal and unrelated,” and “fluent but inadequate.” [4] Whereas in the

context of text summarization, AI hallucinations refer to generated content that is "inconsistent with the source document." [5]

The term has definitely captured the popular imagination. "Hallucinate" secured its position as the word of 2023 [6] and Dictionary.com noted a 46% surge in searches for the term over the past year. [7] The popular press has also adopted this usage as reflected in hundreds of news articles on the subject.

For the sake of our discussion, a working definition of a hallucination we'll use is this: AI generated output that contains false information.

Schwartz Case

The first and most famous of all AI hallucinations happened in the case of *Mata v. Avianca*. [8] Mata's attorney, Steven A. Schwartz, famously used ChatGPT for legal research and filed a brief in federal court that contained the fruit of that research: fake cases with fake citations and fake holdings. Schwartz has been publicly ridiculed and was ultimately sanctioned by the court for failing to check the output and for attempting to cover up the failing. [9] Schwartz remarked, " I just never thought [the cases] could be made up." Peter LoDuca, the attorney of record associated with Schwartz, testified that it "never crossed [his] mind" that the cases were bogus. [10]

Why Do Hallucinations Happen?

So why do LLMs produce false information and how can this result from their *normal* operation? To understand the why, we must first explore the how: How do LLMs operate? How do they create output?

Stochastic Parrot

At this point, you've probably heard AI experts talk about generative AI as a "language calculator" or a "stochastic parrot." They describe, at a high level, that the LLM functions like a language calculator, determining what the next most probable word would be based upon your input. So, for example, if you inputted, "A cow goes..," the LLM would complete the sentence with "moo" because that sound lives in the same

constellation of meaning. I myself have used this astrophysical metaphor, but it overgeneralizes what's happening under the hood.

A Web of Meaning

Large language models, as their name implies, are trained on large amounts of data. GPT-4 was trained on a 13 trillion token dataset of text and code-based data, resulting in a model size of 1.8 trillion parameters, which was 10 times larger than GPT-3. The purpose of this training is to imbue the model with the ability to understand language, the meaning of words and their relationship to other words, context, differences in format (for example, news versus screenplays), intent, dialect and jargon, and more.

What is absent in this training is a grounding in truth. LLMs generate output based on the probabilistic relationships learned from the training data. Like a web, what is relevant to the LLM is the interconnectedness of one word to another word. This relationship is relative and its absolute truth or falsity is irrelevant. This is because the highest probability next token is not necessarily the one that is factually accurate.

Garbage In, Garbage Out

That said, we cannot ignore the obvious: many times the Generative AI output *is* factually accurate. This result has less to do with the fact of its truth. Rather, the factually correct output is generated because factually correct source data was used for training and that same data was relevant to the user's request. Or, to put it more succinctly: garbage in, garbage out. If high quality, factually accurate information is fed into the system, and then called upon, then it is much more likely that the LLM will output high quality and factually correct information.

Attention is All You Need

Let's get a little more technical about this interconnected "web of meaning" I referenced earlier. You may have heard about transformers, the innovation underlying the current upward surge in AI development. Introduced in the landmark paper "Attention is All You Need," [11] the transformer architecture brought a fundamental shift in how language models process and understand text.

Before transformers, recurrent neural networks (RNNs) were the go-to architecture. RNNs processed text sequentially, one word at a time, maintaining hidden states to capture previous context. Despite their ability to handle long sequences, RNNs were inherently limited by their sequential nature. This led to slow training and a loss of meaning for words that may be relevant to each other but were out of sequence.

Transformers, however, revolutionized this process through the concept of self-attention. Instead of processing tokens one by one, the self-attention mechanism allows each token in the sequence to directly relate to every other token at the same time. Each token receives attention scores for every other token, computing a weighted sum of these scores to form a context-aware representation. This enables the model to capture interconnected meaning in a single forward pass.

The ability to capture broad contextual relationships through self-attention helps transformers generalize well across different tasks. This flexibility improves their performance not only in next-word prediction but also in other language understanding tasks. This is why transformers have become the backbone of today's most advanced large language models.

How to Fix Hallucinations?

Based on the discussion so far, you would be right to be suspicious of the idea of eliminating hallucinations entirely. Certainly, they cannot be “fixed” in the sense of eliminating the statistical probability engine that lies at the very heart of large language models, including next word prediction. That said, it is possible to employ tactics to minimize factual errors. So, how do we minimize these pesky missteps?

Here are a few ways:

Curated Datasets

Carefully selecting training datasets that prioritize factual accuracy can significantly reduce hallucinations. While the sheer volume of training data ensures broad generalization, it’s crucial to ensure this data includes vetted and reputable sources. By feeding the models a diet of reliable information, we increase the likelihood that their

outputs will be grounded in truth.This approach lays a foundation where the relationships and patterns formed are more aligned with reality.

Reinforcement Learning from Human Feedback (RLHF)

Another layer of quality control comes from RLHF or a “human in the loop,” where human evaluators rank factual correctness as well as other factors of model-generated responses. This feedback loop fine-tunes the LLM to disfavor toxicity and bias and to favor human values as well as accurate and contextually relevant answers, providing a practical check against erroneous outputs. ChatGPT benefited from RLHF and that is one of the reasons credited for its human and fluent affect.

Retrieval-Augmented Generation (RAG)

Normally, an LLM relies on its parametric memory, which is knowledge stored in its parameters from the data it was trained upon. And, although that training data may be vast, it is not complete: it does not include all of humanity’s knowledge. RAG is a method for infusing an LLM with curated, relevant information by incorporating external databases, search engines, or knowledge graphs to ground the generated responses in factual information. This retrieval-augmented strategy combines the creative, generative power of an LLM with real-time, accurate information retrieval. In this way, RAG closes the gaps of missing, relevant information.

Prompt Engineering and Instruction Tuning

Well-crafted prompts reduce ambiguity and guide the model to more accurate answers. A prompt can be phrased to restrict the model to only provide an answer from the context it is provided and nothing else. Here’s an example:

> `Answer the question as truthfully as possible by only using the provided context. If you cannot find the answer in the context, respond, “I don’t know.”`

Although this instruction is helpful to mitigate the model reaching beyond the context, it alone is not foolproof. Similarly, instruction tuning, where the LLM is trained to follow

specific guidelines that emphasize factual consistency, can also help reduce the frequency of hallucinations.

Domain Specialized Fine-Tuned Models

Fine-tuning models on domain-specific, curated datasets also ensures more accurate results in specific areas like law, medicine, and finance. This targeted training minimizes factual errors when handling complex or high-stakes topics. The reason a fine-tuned model minimizes factual errors is because we are essentially flooding the field with enormous amounts of highly relevant information, like case law, so the model has more accurate data to draw from. And this data, unlike with RAG, gets baked into the model's parametric memory so it is not limited by the size of the model's context window.

Multi-Model Verification

Using an ensemble of models to cross-check responses increases the likelihood of accurate results. An example of this is OpenAI's moderation API, which is called upon to block potentially biased or toxic responses before they are presented to the user. In a similar way, multiple models, with differing domains of knowledge, could be called upon to fact check a response before it is presented to the user. Each model's different perspective would create a consensus that mitigates hallucination risks.

Ultimately, "fixing" hallucinations requires a multi-pronged approach, from enhancing the quality of training data to integrating retrieval systems and refining prompts. It's an evolving field, and as we get better at aligning AI with factual truths, the more we can trust digital assistants in our daily practices. But, for the time being, the best approach is one that always works: "Trust but verify."

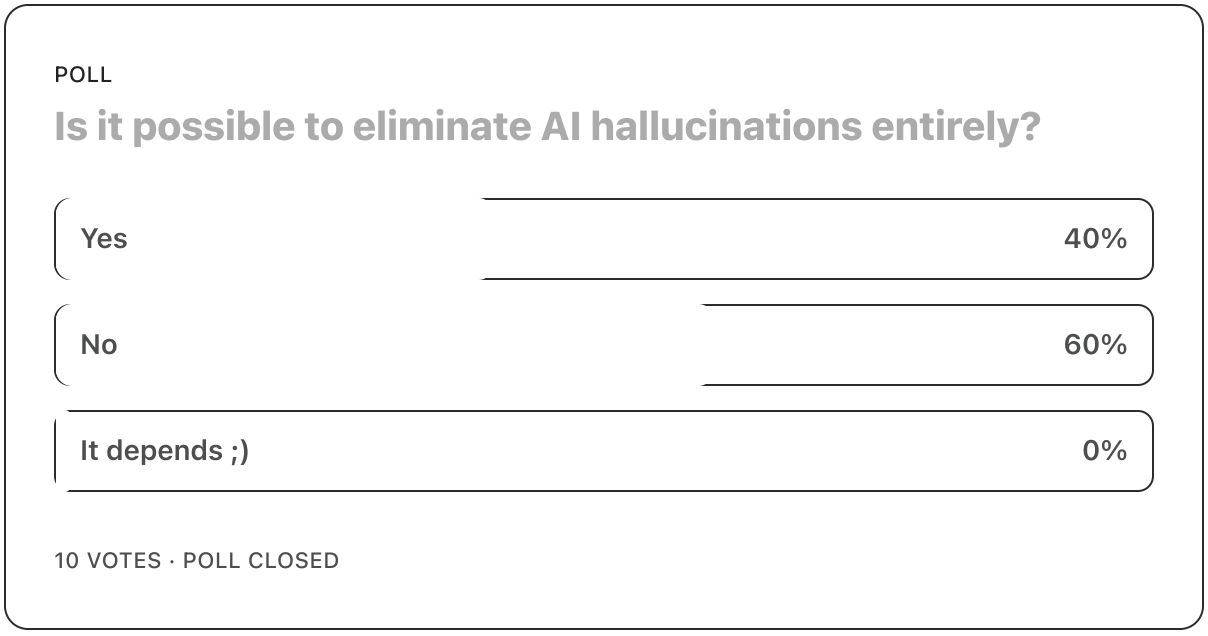
Closing Thoughts

We've come a long way in understanding the nature of hallucinations in large language models (LLMs), dispelling some myths while acknowledging their challenges. Just as Bugs Bunny turns into a mirage-inspired feast in "Wackiki Wabbit," LLMs can sometimes conjure up their own realities—however vivid and convincing they may appear.

Ultimately, AI hallucinations aren't inherently problematic; they simply reflect the current limitations of a technology still finding its feet. With carefully curated datasets, reinforcement learning, and retrieval-augmented generation, we're on the right path toward minimizing these hiccups. The key lies in striking a balance between the creative potential of generative AI and the precision required for specific domains like law, healthcare, and finance.

As we continue to refine these tools, we'll increasingly be able to trust AI to complement and enhance our work without leading us astray. So, next time your virtual assistant offers a curious response, remember that while the hallucination might not be accurate, it's part of a fascinating journey in the ever-evolving world of artificial intelligence.

Stay curious, stay thoughtful, and keep exploring—there's a whole world of knowledge out there waiting to be unlocked!



By the way, if you'd like to learn more about how how AI works and how it will impact the legal profession, you should apply to LawDroid University! It includes all 7 sessions from LawDroid's spectacularly successful AI Conference.

LawDroid University is available for free for everyone to use.

- **Free to use** - It's 100% free educational content for everyone, just sign up below.
- **Insightful** - Get educated about the intersection of artificial intelligence and the law as taught by experts.
- **Value Packed** - Filled with videos, summaries, key takeaways, quotable quotes, transcripts and more! Find sessions on AI and the State of the Art, Ethics, Access to Justice, Practice of Law, Education, and the Business of Law.
  AI Q&A - Ask a chatbot questions about the content and get fully informed answers immediately.

👉 To immerse yourself in this enriching educational voyage, learn more, or sign up, please visit https://lawdroid.com/subscriptions/lawdroid-university/.

---

1 Hallucination, Cambridge Dictionary, https://dictionary.cambridge.org/dictionary/english/hallucination

2 What are AI hallucinations?, Google Cloud, https://cloud.google.com/discover/what-are-ai-hallucinations

3 AI Hallucinations: A Misnomer Worth Clarifying, arXiv:2401.06796v1 [cs.CL] 09 Jan 2024, https://arxiv.org/pdf/2401.06796v1

4 Id.

5 Id.

6 Hallucinate, the Cambridge Dictionary Word of the Year 2023, https://dictionary.cambridge.org/editorial/woty

7 The Dictionary.com Word of the Year is hallucinate, https://content.dictionary.com/word-of-the-year-2023/

8 Here's What Happens When Your Lawyer Uses ChatGPT, New York Times, May 27, 2023, https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

9 Sanctions Order, Mata v. Avianca Inc., June 22, 2023, 22-cv-1461 (PKC), United States District Court, S.D. New York.

10 Id.

11 Attention is All You Need, arXiv:1706.03762 [cs.CL] 12 Jun 2017 (v1), https://arxiv.org/pdf/1706.03762

Comments