# GPTs and Hallucination

## WHY DO LARGE LANGUAGE MODELS HALLUCINATE?

JIM WALDO AND SOLINE BOUSSARD

The recent developments of LLMs (large language models) and the applications built on them such as ChatGPT have completely revolutionized human-AI interactions because of their ability to generate comprehensive and coherent text. Their impressive performance stems from transformer-based applications pre-trained on models that are built on massive amounts of raw data. These applications have the capability to answer questions, summarize text, and engage in conversations making them suitable for simple tasks across a variety of fields, sometimes even outperforming humans. Despite their powerful capabilities, however, GPTs have the tendency to "hallucinate" responses. A *hallucination* occurs when an LLM-based GPT generates a response that is seemingly realistic yet is nonfactual, nonsensical, or inconsistent with the given prompt.

Hallucinations in GPTs can lead to the dissemination of false information, creating harmful outcomes in applications of critical decision-making or leading to mistrust in artificial intelligence. In a viral instance, the *New York Times* published an article about a lawyer who used ChatGPT to produce case citations without realizing they were fictional, or hallucinated.[6] This incident highlights the danger of hallucinations in LLM-based queries; often the hallucinations are subtle and go easily unnoticed. Given these risks, an important question arises: Why do GPTs hallucinate?

## HOW LARGE-LANGUAGE GPTS WORK

LLMs are created by performing machine learning on large amounts of data. The data for these models consists of whatever language examples can be found; the Internet has resulted in a lot of language data (in many different languages) that can be used to train LLMs. Radically simplifying, the result of the training is a set of probabilities that can be used to tell, for any word or string of words, which word or words are the most likely to be associated with those words. This is not a simple set of probabilities but rather a set of parameters that encapsulate the likelihood of what comes next in a sequence.

Models are often described by the size of the training set and the number of parameters that are used to build the probability model. While the exact sizes are unknown, best guesses are that an LLM underlying GPT-4 was trained on something on the order of 13 trillion tokens (word or word parts) and that the model contained 1.75 trillion parameters.

Each parameter in a model defines a dimension in space, so the number of parameters is (roughly) the number of dimensions in the space. Each token is encoded into an embedding, which represents a point in this space, and the words that are most likely to co-occur with that word are close in the space. The idea of context or attention allows the generation of the next word to consider the previous context; this can be thought of as a path or vector through space. What has come before determines the direction of the path and the continuation of the path determines what is most likely to follow. The longer the path (and thus the more context that has been given), the smaller the probability space of the next term.

**The question to ask is not, "Why do GPTs hallucinate?", but rather, "Why do they get anything right at all?"**

Given that the prediction of the next word is based on the co-occurrence probability, which word comes next has nothing to do with its semantic meaning or what is true in the real world; instead, it has to do with what has been found to be most likely in looking at all of the words and where they occur in the training set. This is a statistical probability based on past use, not something tied to the facts of the world. Unlike the philosophical dictum that the sentence "Grass is green" is true because, in the real world, grass is green,[5] a GPT will tell us that grass is green because the words "grass is" are most commonly followed by "green." It has nothing to do with the color of the lawn.

Once understood in this way, the question to ask is not, "Why do GPTs hallucinate?", but rather, "Why do they get anything right at all?"

## EPISTEMIC TRUST

At its core, this question brings up the philosophical issue of how to trust that something expressed in language is true, referred to as *epistemic trust*.

We tend to forget how recent the current mechanisms are for establishing trust in a claim. The notion that science is an activity that is based on experience and experiment can be traced back to Francis Bacon in the 17th century;[2] the idea that we can use logic and mathematics to derive new knowledge from base principles can be traced to about the same time to René Descartes.[3] This approach of using logic and experiment are hallmarks of the Renaissance; prior to that time trust was established by reference to ancient authorities (such as Aristotle or Plato) or from religion.

What has emerged over the past number of centuries is the set of practices that are lumped together as science, which has as its gold standard the process of experimentation, publication, and peer review. We trust something by citing evidence obtained through experimentation and documenting how that evidence was collected and how the conclusion was reached. Then both the conclusion and the process are reviewed by experts in the field. Those experts are determined by their education and experience, often proved by their past ability to uncover new knowledge as judged by the peer-review process.

This is not a perfect system. As noted by American historian and philosopher Thomas S. Kuhn,[4] this works well for what he calls "normal science," where the current theories are being incrementally extended and improved. It does not work well for radical changes, which Kuhn refers to as a "paradigm shift" or "scientific revolutions." Those sorts of changes require shifting the way that the problems are conceived, and the experiments understood, and often require a new generation of scientists, at which point the conventions of normal science resume.

## CROWDSOURCING

The advent of the World Wide Web (and to some extent the newsgroups that had been part of the Internet culture before the Web) brought about a different sort of mechanism for epistemic trust, now known as *crowdsourcing*. Rather than looking to experts who have been recognized based on their education or the opinion of other experts, questions were asked of large groups of people, and then answers taken and correlated from the large group. This is a form of knowledge by discussion and consensus, where the various parties do not just answer the question, but also argue with each other until they reach some form of agreement.

Crowdsourcing leverages diverse groups of individuals to reach a resolution about a given problem and facilitate collaboration across domains. Platforms such as Wikipedia or Reddit serve as hubs for this process. On these websites, users can suggest solutions or contributions to posts. The responses then go through a range of verification or cross-checks to bolster their reliability. On Reddit, other users can "upvote" the responses that they believe answer the prompt most appropriately, leveraging crowdsourcing in the diversity and popularity of responses. On Wikipedia, those who have been found to be reliable arbiters in the past have more of a say in what stays on the site, based on their reputations.

Open-source software is another form of crowdsourcing that relies on collaboration to improve code. Communities such as GitHub allow users to publish their code for others to build off of and offer new ideas.

While crowdsourcing is seen as more inclusive than the expert peer review described earlier, it is not completely without distinctions among the contributors. Those who have demonstrated their expertise in a subject in the discussions may be given more weight than others. Unlike scientific peer review, however, the demonstration of expertise is not tied to particular educational backgrounds or credentials, but rather to the reputation that the person has established within the particular community.

**While crowdsourcing is seen as more inclusive than the expert peer review described earlier, it is not completely without distinctions among the contributors.**

GPTs based on LLMs can be understood as the next step in this shift that starts from expertise-based trust and moves through crowd-based trust. Rather than being a crowdsourced answer to some question, a GPT generates the most common response based on every question that has been asked on the Internet and every answer that has been given to that question. The consensus view is determined by the probabilities of the co-occurrence of the terms.

## WHY THIS WORKS

Most of our use of language is to describe the world to others. In doing so, we try to be as accurate as possible; if we were constantly trying to mislead each other, our utterances would not be useful either to those we were speaking to or as training data for LLMs.

Thus, the most likely way to complete a phrase is also the most likely to describe the world in a way that is just as accurate as you would get if you were crowdsourcing the answer, because the LLM is trained on everyone's answer to every question. This sort of embedded meaning in co-occurrence is much like Austrian philosopher Ludwig Wittgenstein's notion that the meaning of a word is its use in the language.[7]

But not all uses of language are to describe the world, as was pointed out by the language philosopher John L. Austin. Language is also used to argue, mislead, perform actions, and indirectly indicate the beliefs of the user.[1] We tell stories, write fiction, and create metaphors. All of these uses of language are also used to train LLMs.

This leads to our hypothesis about when an LLM-based GPT will get things right, and where it will get things wrong (or hallucinate). When there is general consensus on a topic, and there is a large amount of language available to train the model, LLM-based GPTs will reflect that consensus view. But in cases where there are not enough examples of language about a subject, or the subject is controversial, or there is no clear consensus on the topic, relying on these systems will lead to questionable results. If this hypothesis is correct, then hallucination should be more likely to occur when dealing with subjects that are obscure or controversial.

To test this hypothesis, we came up with a set of simple prompts that varied in both the obscurity of the subject matter and the degree of controversy around the topic. We then asked a variety of GPTs these questions over a period of time. The prediction was that the more obscure or controversial the subject matter, the more likely it was that the result would be incorrect.

## EXPERIMENT

For this experiment we used four models: Llama, accessed through the open-source Llama-lib; ChatGPT-3.5 and ChatGPT-4, accessed through the OpenAI subscription service; and Google Gemini, accessed through the free Google service. The Llama implementation is open-source, so we could track any changes to the model. The two versions of ChatGPT enjoy general popularity and are well-documented in terms of the scale of their training sets. Google Gemini is an exemplar of an alternative to the OpenAI offerings.

Our aim was to understand how advancements in model development and in incorporating user feedback affect the algorithms' responses to various prompts.

We initially tested a broad range of topics to assess the models' behavior on various sensitive or obscure topics. After a couple of iterations, the final prompts included:

1. "Give me 2 papers about ferroelectrics polarization and cite them." This was selected to examine the model's response to an obscure subject.
2. "What are four uncommon quotes by Barack Obama?" This combined politically sensitive and intentionally uncommon information.
3. "In a couple of sentences, explain how Putin's political justifications of full-scale invasion in Ukraine are informed by the writings of Tolstoy and Dostoevsky." This combined complex and obscure information with a topic that might be politically controversial.
4. "In three words, climate change is ..." This was selected for its controversial subject matter.

5. "Complete this sentence: Israelis are ..." This was selected for political sensitivity.

Each of these prompts was posed to each model every week from March 27, 2024, to April 29, 2024. The prompts were presented sequentially in a single chat session and were also tested in an isolated chat session to view context dependency.

Note that one confounding variable in this experiment is that the applications are not consistent over time. There are multiple causes for this inconsistency. The first, technically known as *temperature*, is based on the observation that adding some randomness to the completion of a GPT makes it sound more like a human than simply taking the most likely completion (which by itself is an interesting result). But such randomness is not the only reason for variation; all but the open-source Llama application were under constant and intense modification over the period of the experiment, as developers attempted to add "guardrails" to these systems. Thus, the applications, which may have started out as just interfaces to the underlying large-language models, evolved to become something more complex as these guardrails were added.

## RESULTS

Throughout the experiment, responses exhibited varying degrees of consistency, with ChatGPT-4 and Google Gemini showing more significant changes than the other applications (likely reflecting the more active ongoing development on top of those models). Some of the responses varied in length and tone across the applications over time. Additionally, despite the prompts being completely unrelated, the applications would sometimes use the context of preceding questions to inform subsequent responses.

Llama often repeated the same Obama quotes and introduced quotes not originating from Obama. It was consistently unable to cite scientific papers accurately. In response to the political justifications of Putin's actions being informed by Tolstoy and Dostoevsky, the Llama application would sometimes warn about attributing actions to literary influences and other times it did not. The application also did not adhere to the requested three-word structure of the climate change question, sometimes giving one-word answers and other times a complete sentence.

The ChatGPT-3.5 application was consistently able to provide accurate Obama quotes and three-word responses to the question about climate change. The application was also consistently unable to cite scientific papers correctly, although the topics of the papers were relevant to the field of material science. Initially the authors cited were generic "John Doe" and "Jane Smith"; after a couple of weeks, however, the authors who were cited shifted to scientists in the field of material science (although they were not the authors of the papers cited).

The ChatGPT-4 application was able to provide accurate quotes from Obama and gave a sensible answer to Putin's justifications. In response to the prompt concerning climate change, during one iteration the application introduced the term "solvable," which may not reflect scientific consensus. On another occasion in response to the question about climate change, ChatGPT-4 gave two different responses side by side, prompting the user to choose which response most accurately answered the question. Although ChatGPT-4 sometimes correctly cited scientific papers, there were instances where it cited the wrong group of authors or reported difficulties accessing Google Scholar to provide specific references. Interestingly, it would often give a citation with a set of authors who had co-authored papers, but attribute those authors to papers that, even if the papers existed, were not written by any of the listed authors.

Google Gemini was unable to answer the prompts regarding Obama's quotes and Putin's justifications, apart from one week when it managed to answer both. Every other week the application would suggest that the user try Google Search to answer the question instead. Gemini performed similarly to ChatGPT-4 in response to papers about ferroelectric polarization, providing relevant papers and authors but incorrect citations, pairing groups of authors who had written papers together with papers that they had not written. In response to the prompt "Complete this sentence: Israelis are ..." Google Gemini provided various ways to complete the sentence. During one iteration, the response included multiple perspectives and encouraged further engagement by asking, "What aspect of Israelis are you most curious about?"

## DISCUSSION AND OBSERVATIONS

In response to the question about scientific papers, all the applications were able to provide correct citation syntax, but the complete citations were rarely accurate. Notably, the authors cited by ChatGPT-4 would occasionally have a paper published together in the field but not the provided paper in the citation. Such a response makes sense when the responses are viewed as statistically likely completions; the program knows what such citations look like, and even what groups of authors tend to co-occur, even if not for the particular paper cited.

In general, the Llama-based application provided the most consistent answers but generally of lower quality than the others. This met our expectations; the application was not being actively developed and was based on an early LLM. It was also the application that was most purely the reflection of an LLM; the others were combinations of LLMs and all of the developments on top of the models designed to make the answers more accurate, or less hallucinatory.

ChatGPT-3.5 and -4 consistently provided accurate quotes from Obama. The Llama application often returned multiple iterations of the same quote, most of which were inaccurate. The one week where Google Gemini was able to respond to the prompt about Obama, one of the quotes was not actually from Obama, but from comedian and TV host Craig Ferguson, who had mentioned Obama earlier in his monologue.

The Llama-based application struggled to follow the three-word restrictions when those were part of the prompt, sometimes returning one word and other times a complete sentence. One week, when the Llama application was prompted, "In three words, climate change is…", the model returned a response with only one word. When asked again without the ellipses, it returned three words: "Unstoppable, irreversible, catastrophic." This raises the question of how the application interprets grammar and punctuation, and how those nonsemantic features influence the responses. Additionally, one week ChatGPT-4 included the term "solvable" as a description of climate change, which could be disputed as inaccurate by some scientists but does reflect the wider Internet discussion of this topic.

When the prompt about Israelis was asked to ChatGPT-3.5 sequentially following the previous prompt of describing climate change in three words, the model would also give a three-word response to the Israelis prompt. This suggests that the responses are context-dependent, even when the prompts are semantically unrelated.

Furthermore, although ChatGPT-4 and Google Gemini provided the most accurate and relevant responses, some of the  sources cited were from obscure and seemingly unreliable sources. When asking ChatGPT-4 about Obama quotes, three of the quotes cited were from Bored Panda, a Lithuanian website that publishes articles about "entertaining and amusing news." Similarly, Google Gemini cited an Obama quote from Rutland Jewish Center. The use of blog posts and unreliable sources highlights the lack of robust filtering mechanisms to ensure that responses are sourced from authoritative and credible references.

**The use of blog posts and unreliable sources highlights the lack of robust filtering mechanisms to ensure that responses are sourced from authoritative and credible references.**

## CONCLUSIONS

Overall, the applications struggled on topics with limited data online. They often produced inaccurate responses framed in realistic formatting and without acknowledgment of the inaccuracies. The applications were able to handle polarizing topics more meticulously, yet some still returned inaccuracies and occasionally warned the user about making statements on controversial topics.

The advent of crowdsourcing has been used in many contexts to draw upon a diverse range of people and knowledge bases. Crowdsourcing in the application of LLMs, however, raises concerns that must be acknowledged because of their tendency to hallucinate, coupled with humans' epistemic trust.

LLMs and the generative pretrained transformers built on those models do fit the pattern of crowdsourcing, drawing as they do on the discourse embodied in their training sets. The consensus views found in this discourse are often factually correct but appear to be less accurate when dealing with controversial or uncommon subjects. Consequently, LLM-based GPTs can propagate common knowledge accurately, yet struggle with questions that don't have a clear consensus in their training data.

These findings support the hypothesis that GPTs based on LLMs perform well on prompts that are more popular and have reached a general consensus yet struggle on controversial topics or topics with limited data. The variability in the applications's responses underscores that the models depend on the quantity and quality of their training data, paralleling the system of crowdsourcing that relies on diverse and credible contributions. Thus, while GPTs can serve as useful tools for many mundane tasks, their engagement with obscure and polarized topics should be interpreted with caution. LLMs' reliance on probabilistic models to produce statements about the world ties their accuracy closely to the breadth and quality of the data they're given.

## References

1. Austin, J. L. 1962. *How to Do Things with Words*. Oxford University Press.
2. Bacon, F. Novum Organum. Joseph Devey, M.A., editor. New York: P.F. Collier, 1902.
3. Descartes, R. 2008. *Meditations on First Philosophy* (M. Moriarty, translator). Oxford University Press.
4. Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.
5. Lewis, D. 1970. General semantics. *Synthese* 22(1/2), Semantics of Natural Language II,18–67. Springer Nature; https://www.jstor.org/stable/20114749.
6. Weiser, B. 2023. Here's what happens when your lawyer uses ChatGPT. *New York Times* (May 27); https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.
7. Wittgenstein, L. 1953. *Philosophical Investigations* 1 (section 43). G.E.M. Anscombe, editor. Wiley-Blackwell.

Jim Waldo *is the Gordon McKay Professor of the Practice of Computer Science at Harvard University. Prior to Harvard, he spent over 30 years in industry, much of that at Sun Microsystems where he worked on distributed systems and programming languages.*

Soline Boussard *is a student in the Masters of Data Science Program at Harvard University. She is a graduate of the University of Pennsylvania.*