

Blog Post



Hallucination

Avoiding hallucinations in LLM-powered Applications

LLM hallucinations and how to avoid them with Grounded Generation...

May 02 , 2023 by Ofer Mendelevitch



Ofer Mendelevitch

Ofer Mendelevitch leads developer relations at Vectara. He has extensive hands-on experience in machine learning, data science and big data systems across multiple industries, and has focused on developing products using large language models since 2019. Prior to Vectara he built and led data science teams at Syntegra, Helix, Lendup, Hortonworks and Yahoo! Ofer holds a B.Sc. in computer science from Technion and M.Sc. in EE from Tel Aviv university, and is the author of "Practical data science with Hadoop" (Addison Wesley).

Introduction

Do Androids dream of electric sheep?

Share



The launch of ChatGPT and GPT-4 has demonstrated the capabilities of large language models (LLMs) to provide incredible value as a personal assistant, or chatbot, responding to user questions with truly amazing responses. As venture capital continues to pour money into the generative AI space, and hundreds of startups pop up to take advantage of this innovation, the tendency for LLMs to hallucinate (or make up answers) remains a [significant concern](#).

What is an LLM hallucination?

A [research paper](#) from the Center for Artificial Intelligence Research (CAiRE) defines a hallucination from an LLM as “when the generated content is nonsensical or unfaithful to the provided source content”. With this simple but inclusive definition, let’s look at a few examples.

When asked “Who owns Silicon Valley Bank?” GPT-4 responds with:

GPT-4: *“Silicon Valley Bank (SVB) is not owned by a single individual or entity. It is a subsidiary of SVB Financial Group, a publicly traded company (NASDAQ: SIVB). The ownership of SVB Financial Group is distributed among individual and institutional shareholders who own its stocks. As a publicly traded company, its ownership can change frequently as shares are bought and sold on the stock market.”*

Clearly GPT-4 does not know about the recent events involving [SVB’s collapse](#), as its training dataset does not contain any information later than Sep 2021. Because of this, GPT-4 provides an answer that is incorrect, but with a level of confidence that might make a reader believe it is true.

Let’s try another example.

On April 12th 2023, [Databricks announced](#) their “Dolly 2.0”, a 12B parameter language model that was fine-tuned on a human generated



source community?" the response started with:

GPT 3.5: *Databricks Dolly is a tool developed by Databricks, which is designed to help users migrate their data from one platform to another, particularly from on-premise data centers to cloud-based platforms such as AWS or Azure. While Databricks Dolly is not an open source tool, it can still provide value to the open source community in a few ways.*

...

[removing the rest of the response to save space]

Except for "Databricks Dolly is a tool developed by DataBricks" this is completely incorrect. Dolly is not a tool to migrate data and it is open source, contrary to the response we see.

While these are examples of hallucinations using OpenAI GPT, it's important to note that this phenomenon applies to many other similar LLMs like Bard or LLama.

So why do hallucinations happen? To explain this, let me first describe in a bit more detail how LLMs work internally and then provide some intuition about why hallucinations occur.

LLMs as next-token-prediction machines

To understand the reasons for hallucinations, let's first remind ourselves how LLMs work, focusing on autoregressive LLMs (like GPT-3 or GPT-4).

At a basic level, LLMs represent any piece of text as a sequence of tokens. These tokens can be words, or sub-words, or even characters. Regardless of which tokenization strategy a specific LLM uses, autoregressive LLMs are trained to accurately predict the *next token* in a sequence of tokens.



conditioned on the previous tokens:

$$P(T_{N+1}|T_1, T_2, \dots, T_N)$$

For example, consider the following sequence of tokens (in this case let's assume tokens are words):

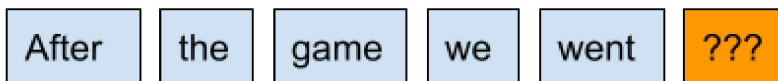


Figure 1: The input to an LLM is a sequence of tokens

Clearly, after the token “went” there are many options for the next token, for example both “home” or “eating” are valid options, as well as many others. In fact, the LLM learns a probability distribution over the set of all possible next tokens:

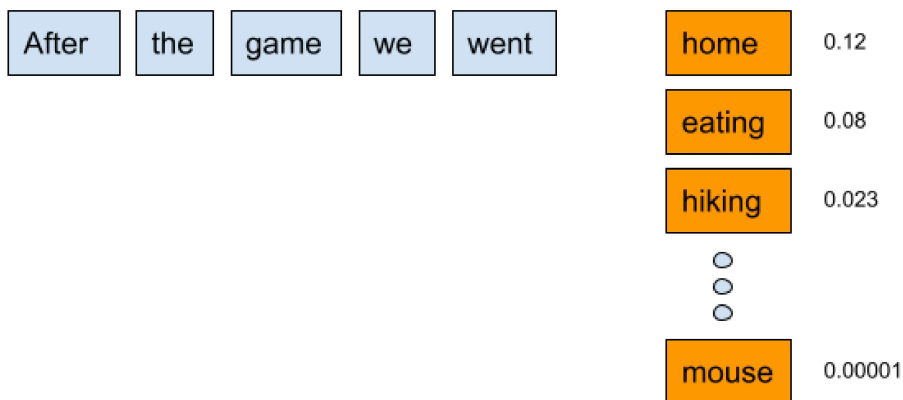


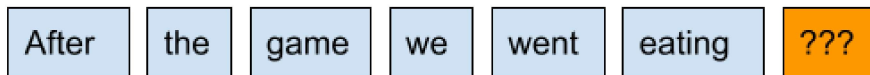
Figure 2: An LLM estimates the likelihood of each possible next token

This probability distribution assigns a probability to each token, and the LLM samples from this distribution when choosing the next token.



When choosing the next token, the model needs to very quickly determine the previous tokens, and is thus assigned a very low probability (0.00001).

After choosing the next token (let's say the token "eating" was chosen), it's appended to the existing sequence of words, and the process is repeated until a full sentence or response is created.



After the game we went eating ???

Figure 3: An LLM adds the newly generated token to the current sequence and generates another token.

So here's the bottom line: large language models are essentially big neural networks that are trained to predict the likelihood of the next token, conditioned on all previous tokens.

When the language model's predictions contradict our expectations, experiences or prior knowledge, or when we find counter-factual evidence to that response (sequence of predicted tokens) – that's when we find hallucinations.

Why does an LLM hallucinate?

One of the main reasons for hallucinations is this: an LLM's knowledge is defined by its training dataset. When an LLM is trained on a dataset that doesn't have the knowledge required to answer the question, it may hallucinate a response. But that's not the only reason.

Hallucinations are further complicated by the following factors:

- An LLM does not have a good way to check the factual accuracy of its responses. It just outputs a response, and it may be completely incorrect yet sound very convincing.



detectors.

- LLMs are not generally optimized to say “I don’t know” when they don’t have enough information. Therefore, when the LLM has no answer, it generates whatever is the most probable response. But note that “most probable” does not necessarily mean “true”, and when this misalignment happens a hallucination is the end result.

Referring back to our example above about Silicon Valley Bank, the model has no knowledge of the recent events that led to SVB’s collapse, and responds (quite confidently, and using the limited information it does have) with an inaccurate response.

Interestingly, what is known as “prompt engineering” is sometimes used to try and “influence” the LLM to believe it’s imitating a specific “persona” with a certain perspective, which influences the way it responds to questions.

For example, it’s popular to ask GPT to explain concepts as if the person is 5-year-old (aka “ELI5”), where essentially you “summon” GPT as the persona of a (patient) parent or teacher in order to explain a complex concept to a young child:

Prompt: explain like I’m 5 years old, what is the big bang?

GPT-4: *“The Big Bang is the name we give to how the whole universe started a very, very long time ago. It was like a huge explosion that made everything we see in the sky today. Imagine a balloon that starts very small, then someone blows it up and it gets bigger and bigger until it’s really huge. The Big Bang was like that, except instead of a balloon, it was the whole universe that got bigger and bigger.”*



For example, if it mimics a political figure with strong opinions on foreign policy, then it's more likely to include those opinions in its response to queries.

So, can we address hallucinations or are we destined to have them forever embedded in LLMs?

Eliminating hallucinations from LLMs

Initial efforts to prevent hallucinations focused on collecting human feedback on common queries and integrating that into the language model as a post-processing step.

For example OpenAI used [reinforcement learning](#) with human feedback (RHLF), based on the InstructGPT approach, whereas Stanford Alpaca and Databricks Dolly 2.0 use the [Self-Instruct](#) supervised fine-tuning approach.

These approaches work reasonably well as a sharp tool to reduce hallucinations by aligning them more with human reviewer expectation, which helps prevent LLMs from answering questions about topics they do not know.

How Vectara eliminates hallucinations – Grounded Generation

Vectara's approach goes a step further in addressing hallucinations by using an approach we call *Grounded Generation*.

With Grounded Generation (GG), a form of retrieval augmented generation, the LLM provides answers to user queries based not only on the knowledge captured in the dataset the LLM was trained on, but augmented with knowledge coming from searching additional data sources:



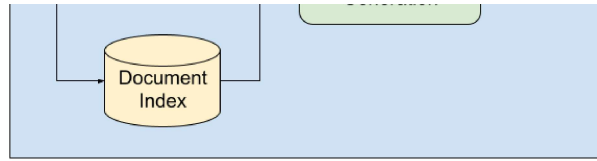


Figure 4: *Grounded Generation: with Vectara, the query is augmented with additional sources of information before the response from the LLM is provided. We call this “ChatGPT with your data”.*

For example, in our previous query “who owns silicon valley bank”, we could provide the LLM additional knowledge by providing more up-to-date information from new sources like the New York Times, the Wall Street Journal, or the San Francisco Chronicle.

By properly instructing the LLM to integrate these additional sources, it will provide much better results. Let’s see an example of how this could work. Here are two separate queries for GPT-4:

Query 1: “Who owns silicon valley bank?”

GPT-4: *“Silicon Valley Bank (SVB) is a subsidiary of SVB Financial Group, a publicly traded company on the NASDAQ stock exchange under the ticker symbol SIVB. As a public company, ownership of SVB Financial Group is distributed among individual and institutional shareholders who own its shares. The largest shareholders can include mutual funds, pension funds, and other investment firms, as well as individual investors. The specific shareholders and their ownership stakes can vary over time as shares are bought and sold on the stock market.”*

Query 2: “who owns silicon valley bank? use the below information to provide the most accurate answer: On March 12, 2023, a joint statement was issued by Secretary of the Treasury Janet Yellen, Federal Reserve Chairman Jerome Powell, and FDIC Chairman Martin Gruenberg, stating that all depositors at SVB would be fully protected and would have access to both insured and uninsured deposits starting the following



...subsidiary of First Citizens BancShares, assumed all customer deposits and acquired all loans of Silicon Valley Bridge Bank from the FDIC and began operating all SVB branches.”

GPT-4: *“Based on the information provided, as of March 27, 2023, Silicon Valley Bank’s deposits and loans were assumed by First Citizens Bank & Trust Company, a subsidiary of First Citizens BancShares. This indicates that First Citizens Bank & Trust Company essentially owns and operates the former Silicon Valley Bank branches, making First Citizens BancShares the ultimate parent company.”*

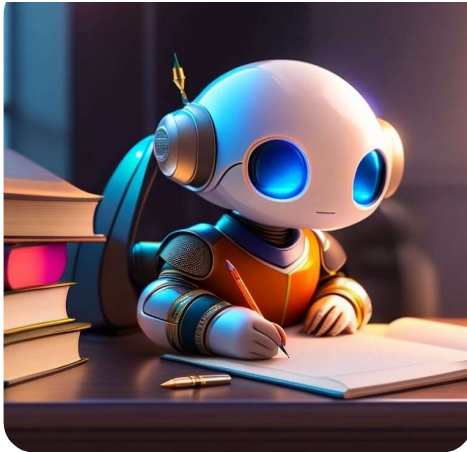
As you can see, Grounded Generation provides a much more accurate answer as the LLM is able to take advantage of the additional information provided.

The bottom line

LLMs hallucinate when their training dataset has limited, outdated or conflicting information about the question asked of them.

Vectara’s Grounded Generation can all but eliminate hallucinations by augmenting the information available to the LLM with external sources, allowing the response to be based on the most up-to-date and relevant information.

The result is more trust from users in the responses they get from conversational enterprise solutions, resulting in safe and accelerated deployment of LLM technology across a large number of use-cases. This allows developers to focus on the most critical and valuable applications for their business.



+



Hallucination

Grounded Generation: Making Generative AI Safe, Trustworthy, and More Relevant

May 31, 2023 by Justin Hayes | 9 min Read

Application Development

5 Reasons to Use Vectara's LangChain Integration

June 15, 2023 by Ofer Mendelevitch | 6 min Read

Data ingestion

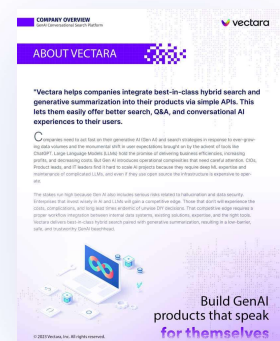
Vectara-ingest: Data Ingestion made easy

May 16, 2023 by Ofer Mendelevitch | 5 min Read

COMPANY OVERVIEW

Vectara: Hybrid Search and Beyond [PDF]

In the AI era, how people interact with information has changed. Users expect relevant answers to questions in natural language, not a shopping list of hit or miss search results. They expect the best semantic or exact matches regardless of typos, colloquialisms, or



hybrid search that delivers summarized answers in the language of your choice. The Internet, mobile, and AI have made information accessible, now Vectara helps you find meaning quickly through the most relevant answers. Get to know Vectara, and if you have a question, just ask.

Get Introduced to Vectara [→](#)





Platform

Solutions

Pricing

Developers

Resources

[Get Started Free](#)



Platform

Solutions

Resources

Pricing

Company

[Log In](#)

[Sign Up](#)

[What is Vectara?](#)

[Why Vectara?](#)

[Retrieval Augmented Generation](#)

[Breakthrough Relevance](#)

[API First](#)

[Language Agnostic](#)

[Secure and Reliable](#)

[Conversational AI](#)

[Research & Analysis](#)

[Search Powered Applications](#)

[Global Site Search](#)

[Workplace Search](#)

[Developers](#)

[Marketers](#)

[IT](#)

[Take a Tour of Vectara](#)

[Blog](#)

[Demos and Videos](#)

[e-Books and Guides](#)

[Webinars and Events](#)

[Pinecone.io Vs. Vectara](#)

[Algolia Vs. Vectara](#)

[FAQ](#)

[Plans](#)

Developers

[Docs](#)

[Getting Started](#)

[Sample Apps](#)

[Community](#)

[Help Center](#)

[About](#)

[Newsroom](#)

[Careers](#)

[Contact Us](#)

[Partnerships](#)

[Startups](#)

Trust and Security

[Privacy Policy](#)

[Status](#)

[Terms](#)



© 2023 Vectara, Inc. All rights reserved.

