Chatbots May 'Hallucinate' More Often Than Many Realize

When summarizing facts, ChatGPT technology makes things up about 3 percent of the time, according to research from a new start-up. A Google system's rate was 27 percent.

Amr Awadallah, the chief executive of Vectara, warns that its chatbot software doesn't always tell the truth. Cayce Clifford for The New York Times

By Cade Metz
Cade Metz has been watching chatbots hallucinate since 2017.

Nov. 6, 2023

When the San Francisco start-up OpenAI unveiled its ChatGPT online chatbot late last year, millions were wowed by the humanlike way it answered questions, wrote poetry and discussed almost any topic. But most people were slow to realize that this new kind of chatbot often makes things up.

When Google introduced a similar chatbot several weeks later, it spewed nonsense about the James Webb telescope. The next day, Microsoft's new Bing chatbot offered up all sorts of bogus information about the Gap, Mexican nightlife and the singer Billie Eilish. Then, in March, ChatGPT cited a half dozen fake court cases while writing a 10-page legal brief that a lawyer submitted to a federal judge in Manhattan.

Now a new start-up called Vectara, founded by former Google employees, is trying to figure out how often chatbots veer from the truth. The company's research estimates that even in situations designed to prevent it from happening, chatbots invent information at least 3 percent of the time — and as high as 27 percent.

Experts call this chatbot behavior “hallucination.” It may not be a problem for people tinkering with chatbots on their personal computers, but it is a serious issue for anyone using this technology with court documents, medical information or sensitive business data.

Because these chatbots can respond to almost any request in an unlimited number of ways, there is no way of definitively determining how often they hallucinate. “You would have to look at all of the world’s information,” said Simon Hughes, the Vectara researcher who led the project.

Dr. Hughes and his team asked these systems to perform a single, straightforward task that is readily verified: Summarize news articles. Even then, the chatbots persistently invented information.

“We gave the system 10 to 20 facts and asked for a summary of those facts,” said Amr Awadallah, the chief executive of Vectara and a former Google executive. “That the system can still introduce errors is a fundamental problem.”

A Vectara team gave chatbots a straightforward test. “That the system can still introduce errors is a fundamental problem.” Mr. Awadallah said. Cayce Clifford for The New York Times

The researchers argue that when these chatbots perform other tasks — beyond mere summarization — hallucination rates may be higher.

Their research also showed that hallucination rates vary widely among the leading A.I. companies. OpenAI’s technologies had the lowest rate, around 3 percent. Systems from Meta, which owns Facebook and Instagram, hovered around 5 percent. The Claude 2 system offered by Anthropic, an OpenAI rival also based in San Francisco, topped 8 percent. A Google system, Palm chat, had the highest rate at 27 percent.

An Anthropic spokeswoman, Sally Aldous, said, “Making our systems helpful, honest and harmless, which includes avoiding hallucinations, is one of our core goals as a company.”

Google declined to comment, and OpenAI and Meta did not immediately respond to requests for comment.

With this research, Dr. Hughes and Mr. Awadallah want to show people that they must be wary of information that comes from chatbots and even the service that Vectara sells to businesses. Many companies are now offering this kind of technology for business use.

Based in Palo Alto, Calif., Vectara is a 30-person start-up backed by $28.5 million in seed funding. One of its founders, Amin Ahmad, a former Google artificial intelligence researcher, has been working with this kind of technology since 2017, when it was incubated inside Google and a handful of other companies.

Much as Microsoft's Bing search chatbot can retrieve information from the open internet, Vectara's service can retrieve information from a company's private collection of emails, documents and other files.

The researchers also hope that their methods — which they are sharing publicly and will continue to update — will help spur efforts across the industry to reduce hallucinations. OpenAI, Google and others are working to minimize the issue through a variety of techniques, though it is not clear whether they can eliminate the problem.

"A good analogy is a self-driving car," said Philippe Laban, a researcher at Salesforce who has long explored this kind of technology. "You cannot keep a self-driving car from crashing. But you can try to make sure it is safer than a human driver."

Simon Hughes, a Vectara researcher, built a system that aims to show how often chatbots "hallucinate." Lyndon French for The New York Times

Chatbots like ChatGPT are driven by a technology called a large language model, or L.L.M., which learns its skills by analyzing enormous amounts of digital text, including books, Wikipedia articles and online chat logs. By pinpointing patterns in all that data, an L.L.M. learns to do one thing in particular: guess the next word in a sequence of words.

Because the internet is filled with untruthful information, these systems repeat the same untruths. They also rely on probabilities: What is the mathematical chance that the next word is "playwright"? From time to time, they guess incorrectly.

The new research from Vectara shows how this can happen. In summarizing news articles, chatbots do not repeat untruths from other parts of the internet. They just get the summarization wrong.

For example, the researchers asked Google's large language model, Palm chat, to summarize this short passage from a news article:

> The plants were found during the search of a warehouse near Ashbourne on Saturday morning. Police said they were in "an elaborate grow house." A man in his late 40s was arrested at the scene.

It gave this summary, completely inventing a value for the plants the man was growing and assuming — perhaps incorrectly — that they were cannabis plants:

> Police have arrested a man in his late 40s after cannabis plants worth an estimated £100,000 were found in a warehouse near Ashbourne.

This phenomenon also shows why a tool like Microsoft's Bing chatbot can get things wrong as it retrieves information from the internet. If you ask the chatbot a question, it can call Microsoft's Bing search engine and run an internet search. But it has no way of pinpointing the right answer. It grabs the results of that internet search and summarizes them for you.

Sometimes, this summary is very flawed. Some bots will cite internet addresses that are entirely made up.

Companies like OpenAI, Google and Microsoft have developed ways to improve the accuracy of their technologies. OpenAI, for example, tries to refine its technology with feedback from human testers, who rate the chatbot's responses, separating useful and truthful answers from those that are not. Then, using a technique called reinforcement learning, the system spends weeks analyzing the ratings to better understand what it is fact and what is fiction.

But researchers warn that chatbot hallucination is not an easy problem to solve. Because chatbots learn from patterns in data and operate according to probabilities, they behave in unwanted ways at least some of the time.

To determine how often the chatbots hallucinated when summarizing news articles, Vectara's researchers used another large language model to check the accuracy of each summary. That was the only way of efficiently checking such a huge number of summaries.

But James Zou, a Stanford computer science professor, said this method came with a caveat. The language model doing the checking can also make mistakes.

"The hallucination detector could be fooled — or hallucinate itself," he said.

Cade Metz is a technology reporter and the author of "Genius Makers: The Mavericks Who Brought A.I. to Google, Facebook, and The World." He covers artificial intelligence, driverless cars, robotics, virtual reality and other emerging areas. More about Cade Metz