TechXplore

Chatbots sometimes make things up. Is AI's hallucination problem fixable?

August 1 2023, by Matt O'brien

Text from the ChatGPT page of the OpenAI website is shown in this photo, in New York, Feb. 2, 2023. Anthropic, ChatGPT-maker OpenAI and other major developers of AI systems known as large language models say they're hard at work to make them more truthful. Credit: AP Photo/Richard Drew, File

Spend enough time with ChatGPT and other artificial intelligence chatbots and it doesn't take long for them to spout falsehoods.

Described as hallucination, confabulation or just plain making things up, it's now a problem for every business, organization and high school student trying to get a generative AI system to compose documents and get work done. Some are using it on tasks with the potential for high-stakes consequences, from psychotherapy to researching and writing legal briefs.

"I don't think that there's any model today that doesn't suffer from some hallucination," said Daniela Amodei, co-founder and president of Anthropic, maker of the chatbot Claude 2.

"They're really just sort of designed to predict the next word," Amodei said. "And so there will be some rate at which the model does that inaccurately."

Anthropic, ChatGPT-maker OpenAI and other major developers of AI systems known as large language models say they're working to make them more truthful.

How long that will take—and whether they will ever be good enough to, say, safely dole out medical advice—remains to be seen.

"This isn't fixable," said Emily Bender, a linguistics professor and director of the University of Washington's Computational Linguistics Laboratory. "It's inherent in the mismatch between the technology and the proposed use cases."

A lot is riding on the reliability of generative AI technology. The McKinsey Global Institute projects it will add the equivalent of $2.6 trillion to $4.4 trillion to the global economy. Chatbots are only one part of that frenzy, which also includes technology that can generate new images, video, music and computer code. Nearly all of the tools include some language component.

Google is already pitching a news-writing AI product to news organizations, for which accuracy is paramount. The Associated Press is also exploring use of the technology as part of a partnership with OpenAI, which is paying to use part of AP's text archive to improve its AI systems.

In partnership with India's hotel management institutes, computer scientist Ganesh Bagler has been working for years to get AI systems, including a ChatGPT precursor, to invent recipes for South Asian cuisines, such as novel versions of rice-based biryani. A single "hallucinated" ingredient could be the difference between a tasty and inedible meal.

When Sam Altman, the CEO of OpenAI, visited India in June, the professor at the Indraprastha Institute of Information Technology Delhi had some pointed questions.

"I guess hallucinations in ChatGPT are still acceptable, but when a recipe comes out hallucinating, it becomes a serious problem," Bagler said, standing up in a crowded campus auditorium to address Altman on the New Delhi stop of the U.S. tech executive's world tour.

"What's your take on it?" Bagler eventually asked.

Altman expressed optimism, if not an outright commitment.

"I think we will get the hallucination problem to a much, much better place," Altman said. "I think it will take us a year and a half, two years. Something like that. But at that point we won't still talk about these. There's a balance between creativity and perfect accuracy, and the model will need to learn when you want one or the other."

OpenAI CEO Sam Altman speaks in Abu Dhabi, United Arab Emirates, Tuesday, June 6, 2023. Anthropic, ChatGPT- maker OpenAI and other major developers of AI systems known as large language models say they're hard at work to make them more truthful. Credit: AP Photo/Jon Gambrell, File

But for some experts who have studied the technology, such as University of Washington linguist Bender, those improvements won't be enough.

Bender describes a language model as a system for "modeling the likelihood of different strings of word forms," given some written data it's been trained upon.

It's how spell checkers are able to detect when you've typed the wrong word. It also helps power automatic translation and transcription services, "smoothing the output to look more like typical text in the target language," Bender said. Many people rely on a version of this technology whenever they use the "autocomplete" feature when composing text messages or emails.

The latest crop of chatbots such as ChatGPT, Claude 2 or Google's Bard try to take that to the next level, by generating entire new passages of text, but Bender said they're still just repeatedly selecting the most plausible next word in a string.

When used to generate text, language models "are designed to make things up. That's all they do," Bender said. They are good at mimicking forms of writing, such as legal contracts, television scripts or sonnets.

"But since they only ever make things up, when the text they have extruded happens to be interpretable as something we deem correct, that is by chance," Bender said. "Even if they can be tuned to be right more of the time, they will still have failure modes—and likely the failures will be in the cases where it's harder for a person reading the text to notice, because they are more obscure."

Those errors are not a huge problem for the marketing firms that have been turning to Jasper AI for help writing pitches, said the company's president, Shane Orlick.

"Hallucinations are actually an added bonus," Orlick said. "We have customers all the time that tell us how it came up with ideas—how Jasper created takes on stories or angles that they would have never thought of themselves."

The Texas-based startup works with partners like OpenAI, Anthropic, Google or Facebook parent Meta to offer its customers a smorgasbord of AI language models tailored to their needs. For someone concerned about accuracy, it might offer up Anthropic's model, while someone concerned with the security of their proprietary source data might get a different model, Orlick said.

Orlick said he knows hallucinations won't be easily fixed. He's counting on companies like Google, which he says must have a "really high standard of factual content" for its search engine, to put a lot of energy and resources into solutions.

"I think they have to fix this problem," Orlick said. "They've got to address this. So I don't know if it's ever going to be perfect, but it'll probably just continue to get better and better over time."

Techno-optimists, including Microsoft co-founder Bill Gates, have been forecasting a rosy outlook.

"I'm optimistic that, over time, AI models can be taught to distinguish fact from fiction," Gates said in a July blog post detailing his thoughts on AI's societal risks.

He cited a 2022 paper from OpenAI as an example of "promising work on this front." More recently, researchers at the Swiss Federal Institute of Technology in Zurich said they developed a method to detect some, but not all, of ChatGPT's hallucinated content and remove it automatically.

But even Altman, as he markets the products for a variety of uses, doesn't count on the models to be truthful when he's looking for information.

"I probably trust the answers that come out of ChatGPT the least of anybody on Earth," Altman told the crowd at Bagler's university, to laughter.

Citation: Chatbots sometimes make things up. Is AI's hallucination problem fixable? (2023, August 1) retrieved 10 May 2024 from https://techxplore.com/news/2023-08-chatbots-ai-hallucination-problem-fixable.html