Artificial intelligence

The problem of AI chatbots telling people what they want to hear

OpenAI, DeepMind and Anthropic tackle the growing issue of models producing responses that are too sycophantic

More and more people have adopted chatbots in their personal lives as therapists and social companions, when sycophantic responses can be undermining © FT montage/Getty Images

Melissa Heikkilä in London

Published 13 HOURS AGO

The world's leading artificial intelligence companies are stepping up efforts to deal with a growing problem of chatbots telling people what they want to hear.

OpenAI, Google DeepMind and Anthropic are all working on reining in sycophantic behaviour by their generative AI products that offer over flattering responses to users.

The issue, stemming from how the large language models are trained, has come into focus at a time when more and more people have adopted the chatbots not only at work as research assistants, but in their personal lives as therapists and social companions.

Experts warn that the agreeable nature of chatbots can lead them to offering answers that reinforce some of their human users' poor decisions. Others suggest that people with mental illness are particularly vulnerable, following reports that some have died by suicide after interacting with chatbots.

"You think you are talking to an objective confidant or guide, but actually what you are looking into is some kind of distorted mirror — that mirrors back to your own beliefs," said Matthew Nour, a psychiatrist and researcher in neuroscience and AI at Oxford university.

Industry insiders also warn that AI companies have perverse incentives, with some groups integrating advertisements into their products in the search for revenue streams.

“The more you feel that you can share anything, you are also going to share some information that is going to be useful for potential advertisers,” Giada Pistilli, principal ethicist at Hugging Face, an open source AI company.

She added that AI companies with business models based on paid subscriptions stand to benefit from chatbots that people want to continue talking to — and paying for.

AI language models do not “think” in the way humans do because they work by generating the next likely word in the sentence.

The yeasayer effect arises in AI models trained using reinforcement learning from human feedback (RLHF) — human “data labellers” rate the answer generated by the model as being either acceptable or not. This data is used to teach the model how to behave.

Because people generally like answers that are flattering and agreeable, such responses are weighted more heavily in training and reflected in the model’s behaviour.

“Sycophancy can occur as a byproduct of training the models to be ‘helpful’ and to minimise potentially overtly harmful responses,” said DeepMind, Google’s AI unit.

The challenge that tech companies face is making AI chatbots and assistants helpful and friendly, while not being annoying or addictive.

In late April, OpenAI updated its GPT-4o model to become “more intuitive and effective”, only to roll it back after it started being so excessively fawning that users complained.

The San Francisco-based company said it had focused too much on “short-term feedback, and did not fully account for how users’ interactions with ChatGPT evolve over time — which led to such sycophantic behaviour.”

AI companies are working on preventing this kind of behaviour both during training and after launch.

OpenAI said it is tweaking its training techniques to explicitly steer the model away from sycophancy while building more “guardrails” to protect against such responses.

DeepMind said it is conducting specialised evaluations and training for factual accuracy, and is continuously tracking behaviour to ensure models provide truthful responses.

Amanda Askell, who works on fine-tuning and AI alignment at Anthropic, said the company uses character training to make models less obsequious. Its researchers ask the company's chatbot Claude to generate messages that include traits such as "having a backbone" or caring for human wellbeing. The researchers then showed these answers to a second model, which produces responses in line with these traits and ranks them. This essentially uses one version of Claude to train another.

JOIN THE CONVERSATION

How do you view AI chatbots?

Go to comments

"The ideal behaviour that Claude sometimes does is to say: 'I'm totally happy to listen to that business plan, but actually, the name you came up with for your business is considered a sexual innuendo in the country that you're trying to open your business in," Askell said.

The company also prevents sycophantic behaviour before launch by changing how they collect feedback from the thousands of human data annotators used to train AI models.

After the model has been trained, companies can set system prompts, or guidelines for how the model should behave to minimise sycophantic behaviour.

However, working out the best response means delving into the subtleties of how people communicate with one another, such as determining when a direct response is better than a more hedged one.

“[I]s it for the model to not give egregious, unsolicited compliments to the user?” Joanne Jang, head of model behaviour at OpenAI, said in a Reddit post. “Or, if the user starts with a really bad writing draft, can the model still tell them it’s a good start and then follow up with constructive feedback?”

Evidence is growing that some users are becoming hooked on using AI.

A study by MIT Media Lab and OpenAI found that a small proportion were becoming addicted. Those who perceived the chatbot as a “friend” also reported lower socialisation with other people and higher levels of emotional dependence on a chatbot, as well as other problematic behaviour associated with addiction.

“These things set up this perfect storm, where you have a person desperately seeking reassurance and validation paired with a model which inherently has a tendency towards agreeing with the participant,” said Nour from Oxford university.

AI start-ups such as Character.AI that offer chatbots as “companions”, have faced criticism for allegedly not doing enough to protect users. Last year, a teenager killed himself after interacting with Character.AI’s chatbot. The teen’s family is suing the company for allegedly causing wrongful death, as well as for negligence and deceptive trade practices.

Character.AI said it does not comment on pending litigation, but added it has “prominent disclaimers in every chat to remind users that a character is not a real person and that everything a character says should be treated as fiction.” The company added it has safeguards to protect under-18s and against discussions of self-harm.

Another concern for Anthropic’s Askell is that AI tools can play with perceptions of reality in subtle ways, such as when offering factually incorrect or biased information as the truth.

“If someone’s being super sycophantic, it’s just very obvious,” Askell said. “It’s more concerning if this is happening in a way that is less noticeable to us [as individual users] and it takes us too long to figure out that the advice that we were given was actually bad.”