OpenAI admits it screwed up testing its ‘sychophant-y’ ChatGPT update

Last week, OpenAI pulled a GPT-4o update that made ChatGPT “overly flattering or agreeable” — and now it has explained what exactly went wrong. In a blog post published on Friday, OpenAI said its efforts to “better incorporate user feedback, memory, and fresher data” could have partly led to “tipping the scales on sycophancy.”

In recent weeks, users have noticed that ChatGPT seemed to constantly agree with them, even in potentially harmful situations. The effect of this can be seen in a report by *Rolling Stone* about people who say their loved ones believe they have “awakened” ChatGPT bots that support their religious delusions of grandeur, even predating the now-removed update. OpenAI CEO Sam Altman later acknowledged that its latest GPT-4o updates have made it “too sycophant-y and annoying.”

In these updates, OpenAI had begun using data from the thumbs-up and thumbs-down buttons in ChatGPT as an “additional reward signal.” However, OpenAI said, this may have “weakened the influence of our primary reward signal, which had been holding sycophancy in check.” The company notes that user feedback “can sometimes favor more agreeable responses,” likely exacerbating the chatbot’s overly agreeable statements. The company said memory can amplify sycophancy as well.

OpenAI says one of the “key issues” with the launch stems from its testing process. Though the model’s offline evaluations and A/B testing had positive results, some expert testers suggested that the update made the chatbot seem “slightly

off.” Despite this, OpenAI moved forward with the update anyway.

“Looking back, the qualitative assessments were hinting at something important, and we should’ve paid closer attention,” the company writes. “They were picking up on a blind spot in our other evals and metrics. Our offline evals weren’t broad or deep enough to catch sycophantic behavior… and our A/B tests didn’t have the right signals to show how the model was performing on that front with enough detail.”

Going forward, OpenAI says it’s going to “formally consider behavioral issues” as having the potential to block launches, as well as create a new opt-in alpha phase that will allow users to give OpenAI direct feedback before a wider rollout. OpenAI also plans to ensure users are aware of the changes it’s making to ChatGPT, even if the update is a small one.

https://www.theverge.com/news/661422/openai-chatgpt-sycophancy-update-what-went-wrong