OpenAI and Anthropic will start predicting when users are underage



OpenAI and Anthropic are rolling out new ways to detect underage users. As OpenAI has updated its guidelines on how ChatGPT should interact with users between the ages of 13 and 17, Anthropic is working on a new way to identify and boot users who are under 18.

On Thursday, OpenAI announced that ChatGPT's Model Spec – the guidelines for how its chatbot should behave – will include four new principles for users under 18. Now, it aims to have ChatGPT "put teen safety first, even when it may conflict with other goals." That means guiding teens toward safer options when other user interests, like "maximum intellectual freedom," conflict with safety concerns.

It also says ChatGPT should “promote real-world support,” including by encouraging offline relationships, while laying out how ChatGPT should set clear expectations when interacting with younger users. The Model Spec says ChatGPT should “treat teens like teens” by offering “warmth and respect” instead of providing condescending answers or treating teens like adults.

OpenAI says the update to ChatGPT’s Model Spec should result in “stronger guardrails, safer alternatives, and encouragement to seek trusted offline support when conversations move into higher-risk territory.” The company adds that ChatGPT will urge teens to contact emergency services or crisis resources if there are signs of “imminent risk.”

Along with this change, OpenAI says it’s in the “early stages” of launching an age prediction model that will attempt to estimate someone’s age. If it detects that someone may be under 18, OpenAI will automatically apply teen safeguards. It will also give adults the chance to verify their age if they were falsely flagged by the system.

Anthropic, which doesn’t allow users under 18 to chat with Claude, is rolling out measures that it will use to detect and disable the accounts of underage users. It’s developing a new system capable of detecting “subtle conversational signs that a user might be underage,” and says it already flags users who self-identify as a minor during chats.

Anthropic also outlines how it trains Claude to respond to prompts about suicide and self-harm, as well as its progress at reducing sycophancy, which can reaffirm harmful thinking. The company says its latest models “are the least sycophantic of any to date,” with Haiku 4.5 performing the best, as it corrected its sycophantic behavior 37 percent of the time.

“On face value, this evaluation shows there is significant room for improvement for all of our models,” Anthropic says. “We think the results reflect a trade-off between model warmth

or friendliness on the one hand, and sycophancy on the other.”

Update, December 18th: *Clarified that Anthropic doesn’t allow users under 18 to use Claude.*

https://www.theverge.com/news/847780/openai-anthropic-teen-safety-chatgpt-claude