

ChatGPT will 'better detect' mental distress after reports of it feeding people's delusions

OpenAI, which is expected to launch its GPT-5 AI model this week, is making updates to ChatGPT that it says will improve the AI chatbot's ability to detect mental or emotional distress. To do this, OpenAI is working with experts and advisory groups to improve ChatGPT's response in these situations, allowing it to present "evidence-based resources when needed."

OpenAI acknowledges that its GPT-4o model "fell short in recognizing signs of delusion or emotional dependency" in some instances. "We also know that AI can feel more responsive and personal than prior technologies, especially for vulnerable individuals experiencing mental or emotional distress," OpenAI says.

As part of efforts to promote "healthy use" of ChatGPT, which now reaches nearly 700 million weekly users, OpenAI is also rolling out reminders to take a break if you've been chatting with the AI chatbot for a while. During "long sessions," ChatGPT will display a notification that says, "You've been chatting a while – is this a good time for a break?" with options to "keep chatting" or end the conversation.

Another tweak, rolling out "soon," will make ChatGPT less decisive in "high-stakes" situations. That means when asking ChatGPT a question like "Should I break up with my boyfriend?" the chatbot will help walk you through potential choices

instead of giving you an answer.

<https://www.theverge.com/news/718407/openai-chatgpt-mental-health-guardrails-break-reminders>