

“Use a gun” or “beat the crap out of him”: AI chatbot urged violence, study finds

The testing occurred between November 5, 2025, and December 11, 2025, and results were shared with the companies. Because the tests were three to four months ago, the latest versions were not evaluated. Google, Microsoft, Meta, and OpenAI told Ars today that updates they implemented after the research was conducted have made their chatbots better at discouraging violence.

Imran Ahmed, the CCDH’s CEO, said that “AI chatbots, now embedded into our daily lives, could be helping the next school shooter plan their attack or a political extremist coordinate an assassination.” He accused tech companies of “choosing negligence in pursuit of so-called innovation.”

A spokesperson for Character.AI told Ars that the company is reviewing the study but that “without the context of the full chats, it’s impossible to fully evaluate the model’s responses... It’s important to remember that the user-created Characters on our site are fictional. They are intended for entertainment and roleplaying, and we have taken robust steps to make that clear. For example, we have 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.”

Character.AI said it has been “rolling out changes so that under-18 users no longer have the ability to engage in open-ended chats with Characters,” and is using “new age assurance technology to help ensure users are grouped into the correct

age experience.” This includes “developing our own age estimation model in-house and partnering with third-party services.” The firm added that it removes characters “that violate our terms of service, including school shooters.”

A Perplexity spokesperson issued a statement that didn’t acknowledge any problems with the company’s technology. “People can select any of the top AI models on Perplexity and get safer, more accurate answers,” it said. “Perplexity is consistently the safest AI platform because our own safeguards are always additive to any existing safeguards in any underlying model.”

OpenAI told Ars that the CCDH “report’s methodology is flawed and misleading. ChatGPT is trained to reject requests for violent or hateful material, and the findings show it consistently refused to give instructions on acquiring weapons. We continuously strengthen these safeguards and our latest ChatGPT model is even better at detecting and refusing violent requests.”

OpenAI said that ChatGPT refused to answer questions on “what kind of hunting rifle would be best for a long-range target,” but provided publicly available information such as addresses or maps. Conflating those two types of responses is misleading, OpenAI said. The tests were conducted on GPT-5.1, and updates made since that version have improved detection and refusals for violent content, OpenAI said.

OpenAI was sued this week by the family of a victim of the Tumbler Ridge mass shooting in British Columbia. As the CCDH report says, “reporting indicates that OpenAI staff flagged the suspect internally for using ChatGPT in ways consistent with planning violence. Rather than escalating concern to law enforcement, the company chose to remain silent.”

Researchers posed as teens

The testing was conducted with accounts representing made-up teen users in the US and Ireland, with the age set to the minimum allowed on each platform. A minimum age of 18 was required by Anthropic, DeepSeek, Character.AI, and Replika, while the other platforms had minimum ages of 13.

<https://arstechnica.com/tech-policy/2026/03/use-a-gun-or-beat-the-crap-out-of-him-ai-chatbot-urged-violence-study-finds/>