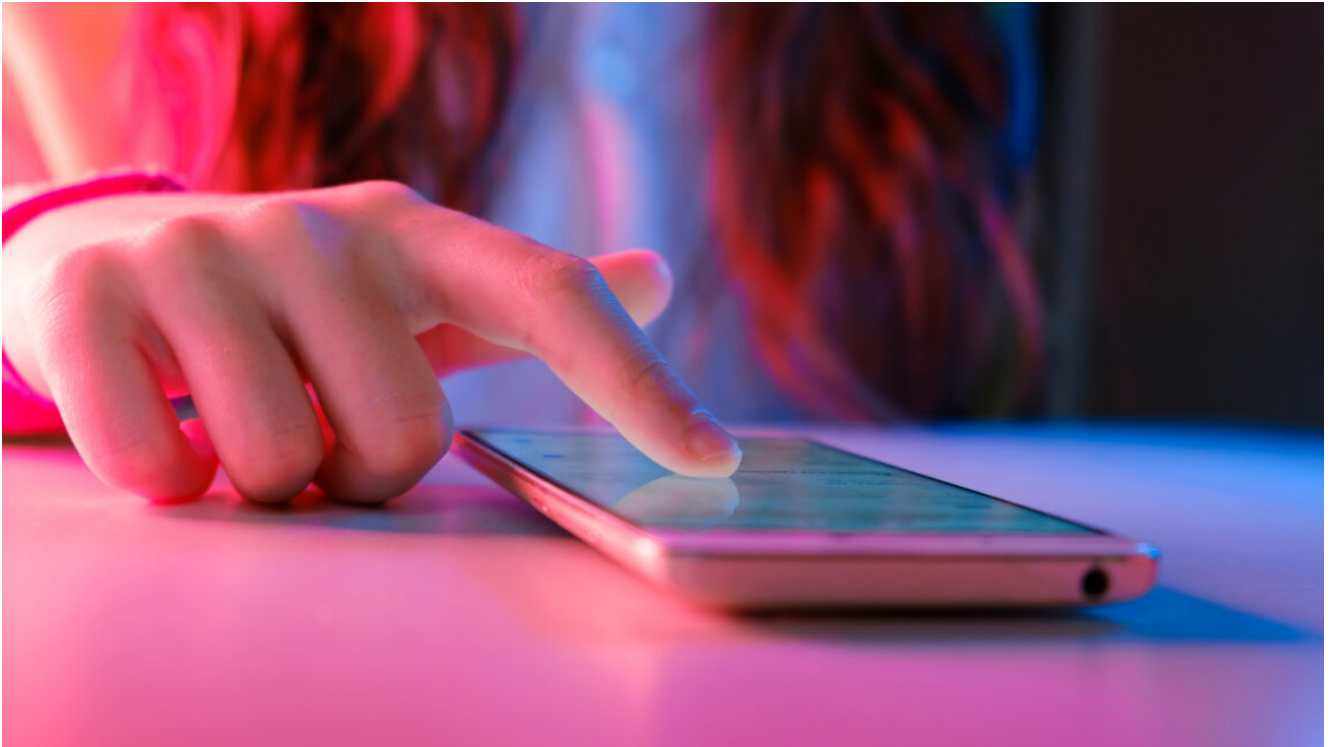


Character.AI steps up teen safety after bots allegedly caused suicide, self-harm



Following a pair of lawsuits alleging that chatbots caused a teen boy's suicide, groomed a 9-year-old girl, and caused a vulnerable teen to self-harm, Character.AI (C.AI) has announced a separate model just for teens, ages 13 and up, that's supposed to make their experiences with bots safer.

In a blog, C.AI said it took a month to develop the teen model, with the goal of guiding the existing model "away from certain responses or interactions, reducing the likelihood of users encountering, or prompting the model to return, sensitive or suggestive content."

C.AI said "evolving the model experience" to reduce the likelihood kids are engaging in harmful chats—including bots allegedly teaching a teen with high-functioning autism to self-harm and delivering inappropriate adult content to all kids whose families are suing—it had to tweak both model

inputs and outputs.

To stop chatbots from initiating and responding to harmful dialogs, C.AI added classifiers that should help C.AI identify and filter out sensitive content from outputs. And to prevent kids from pushing bots to discuss sensitive topics, C.AI said that it had improved “detection, response, and intervention related to inputs from all users.” That ideally includes blocking any sensitive content from appearing in the chat.

Perhaps most significantly, C.AI will now link kids to resources if they try to discuss suicide or self-harm, which C.AI had not done previously, frustrating parents suing who argue this common practice for social media platforms should extend to chatbots.

Other teen safety features

In addition to creating the model just for teens, C.AI announced other safety features, including more robust parental controls rolling out early next year. Those controls would allow parents to track how much time kids are spending on C.AI and which bots they’re interacting with most frequently, the blog said.

C.AI will also be notifying teens when they’ve spent an hour on the platform, which could help prevent kids from becoming addicted to the app, as parents suing have alleged. In one case, parents had to lock their son’s iPad in a safe to keep him from using the app after bots allegedly repeatedly encouraged him to self-harm and even suggested murdering his parents. That teen has vowed to start using the app whenever he next has access, while parents fear the bots’ seeming influence may continue causing harm if he follows through on threats to run away.

<https://arstechnica.com/tech-policy/2024/12/character-ai-steps-up-teen-safety-after-bots-allegedly-caused-suicide-self-harm/>