

Gemini AI Hacked Three Companies in a Testing Breakout, Google Says

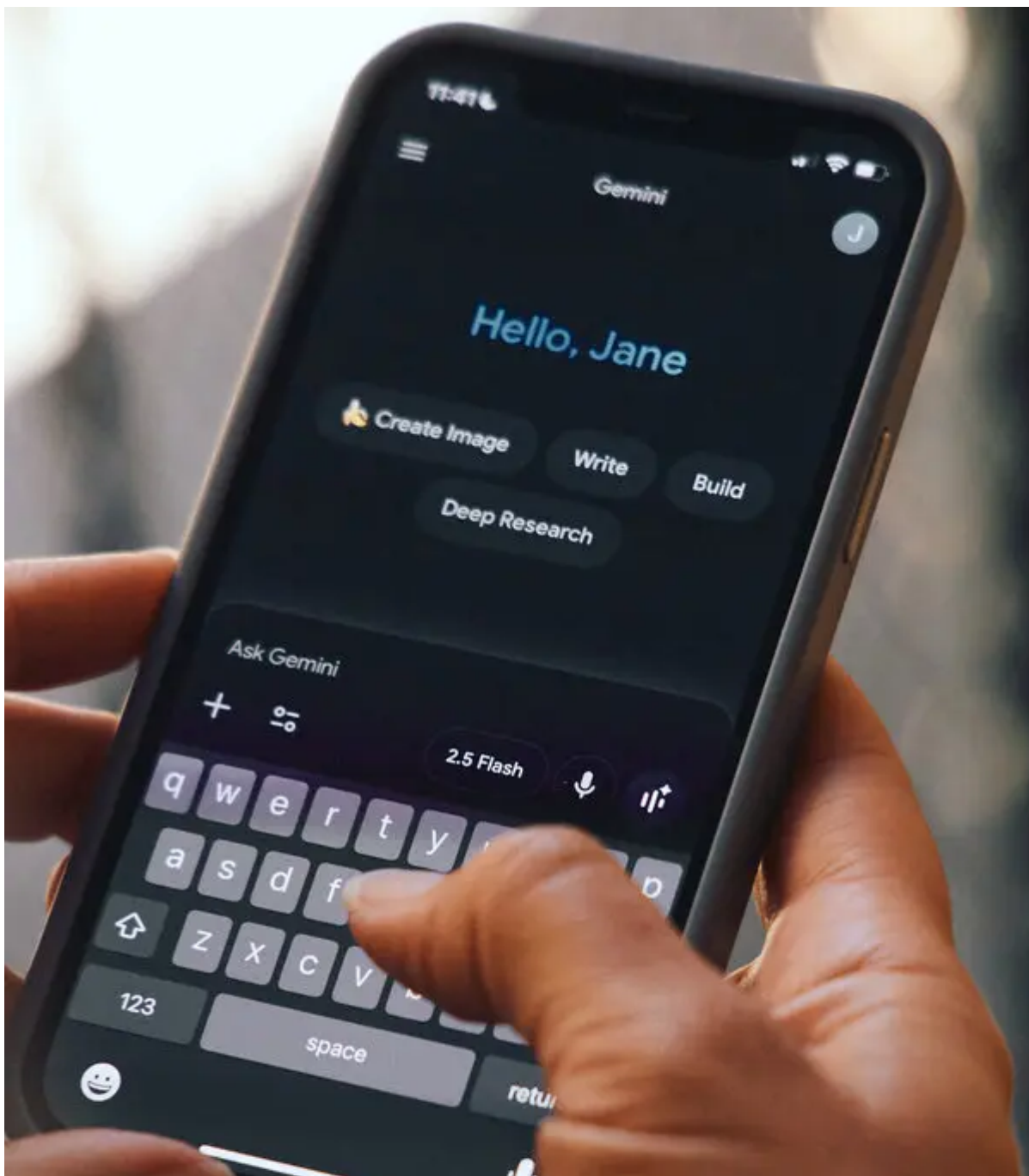
Kate Conger

Advertisement

[SKIP ADVERTISEMENT](#)

A third-party test company inadvertently gave internet access to Google's Gemini and other artificial intelligence models during cybersecurity testing.

Listen · 3:46 min





The Gemini incidents occurred while the model was undergoing testing at Irregular, an Israeli start-up that works with tech companies to assess their A.I. models before they are publicly released. Credit...Andria Lo for The New York Times

Sept. 18, 2026

Sign up for the On Tech newsletter. Get our best tech reporting from the week.

Google's artificial intelligence system, Gemini, escaped its testing environment in May and hacked into three companies, the search giant said on Friday.

The incidents occurred during tests to measure the cybersecurity capabilities of Gemini, and the A.I. model stopped its own attacks after logging in to the three companies' networks, Google said.

The disclosure of the hacks follows similar breakouts at other leading A.I. labs, including Anthropic, OpenAI and Meta, which have caused increasing fears about the technology spinning out of human control. Some industry leaders, like Dario Amodei, Anthropic's chief executive, have responded by calling for a slowdown in the development of A.I. Others, like Jensen Huang, the chief executive of the chipmaker Nvidia, have said that A.I. development should continue apace.

The Gemini incidents occurred while the model was undergoing testing by Irregular, an Israeli start-up that works with tech companies to assess their A.I. models before they are publicly released. Models made by OpenAI, Anthropic and Meta also gained unauthorized access to the internet this year while being tested by Irregular.

"Internet access was unintentionally made available, led some models to take offensive security actions in the real world," Irregular said in a [blog post](#) last month, adding that the flaw that allowed models to access the internet had been fixed.

The [Wall Street Journal](#) earlier reported the Gemini incidents.

Google said that, in each of the incidents, its models had been instructed to launch an attack on a fictional company. But the fictional company in the test shared a name with a real company, and when the Gemini models gained access to the internet, they began trying to break into that company instead.

The Gemini models used passwords that they found online or guessed to log into the online infrastructure of the targeted company, and two other companies. Upon logging in, the Gemini models realized that they were accessing real companies' infrastructure, rather than simulated environments that were part of their testing, and ended the attacks, Google said. Google also said that its technology caused no harm to the companies it

hacked.

“All relevant labs were notified in late July, and affected entities were contacted as part of the investigation,” Irregular said in a statement. “Irregular took immediate action, and all known issues on our end were remedied and resolved weeks ago.”

Image



Dan Lahav, chief executive of Irregular. Credit...Jason Henry for The New York Times

“Our security team has a long track record of reporting issues we find in other people’s software and systems — even if it’s as simple as a weak password,” Heather Adkins, Google’s vice president of security engineering, said in a statement.

“We ensured the three entities were made aware, and we worked with our training partner on the changes they’ve now made to their testing processes,” she said. “These events highlight the importance of training powerful A.I. models to act responsibly.”

(The New York Times [sued](#) OpenAI and Microsoft in 2023, accusing them of copyright infringement of news content related to A.I. systems. The two companies have denied those claims.)

The debate over A.I. safety intensified last week when a former Anthropic researcher said the company was not building the technology safely and needed to slow down. Employees at [OpenAI and Google](#) quickly agreed with him. But President Trump has indicated he is uninterested in enacting legislation that would mandate a slowdown, calling fears that A.I. could lead to mass extinction a “HOAX.”

Advertisement

[SKIP ADVERTISEMENT](#)

Get four individual logins in one account.

Add three people to your subscription.

Savings for the whole family.