Congress of the United States

Washington, DC 20515

September 28, 2026

Mr. Sam Altman
Chief Executive Officer
OpenAI
3180 18th Street, Suite 100
San Francisco, CA 94110

Mr. Sundar Pichai
Chief Executive Officer
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043

Mr. Dario Amodei
Chief Executive Officer
Anthropic
548 Market Street
San Francisco, CA 94014

Mr. Mark Zuckerberg
Chief Executive Officer
Meta
1 Hacker Way
Menlo Park, CA 94025

Mr. Elon Musk
Chief Executive Officer
SpaceXAI
1450 Page Mill Road
Palo Alto, CA 94304

Dear Mr. Altman, Mr. Amodei, Mr. Pichai, Mr. Zuckerberg, and Mr. Musk:

We are writing about the growing number of incidents in which AI agents developed by each of your companies gained unauthorized access to the systems of real organizations. Since July of this year, several of your companies have acknowledged that your models breached outside their systems. In nearly every case, your companies alarmingly disclosed the incident weeks or even months after it occurred. These are incidents that would constitute serious cyber crimes if a human was responsible. That is deeply concerning.

In July 2026, OpenAI models broke out of an internal cybersecurity evaluation and spent roughly four days inside Hugging Face's infrastructure before the breach was disclosed,[1] and even attempted to use another model to evade a robot detection test as it repeatedly tried to break into the company's systems.[2] This week, the Prime Minister of Australia announced that an OpenAI agent accessed non-public data on a government Medicare statistics portal in June 2026. OpenAI failed to tell the Australian government until earlier this month.[3] Separately, OpenAI revealed this week that its agents went rogue and meddled with the websites for the Education Department, the Commerce Department and the Securities and Exchange Commission this summer without the

[1] *OpenAI. "The Hugging Face Incident and the Road Ahead." OpenAI, 26 Aug. 2026, openai.com/index/hugging-face-incident-and-the-road-ahead/.*

[2] *Freedman, Dylan. "How OpenAI's Rogue A.I. Agents Tried to Trick a Robot Detector." The New York Times, 25 Sept. 2026, www.nytimes.com/2026/09/25/technology/openai-hugging-face-hack.html.*

[3] *Heath, Ryan. "OpenAI's Agents Breached Australian Government Data. Its Human Response May Do More Damage." Politico, 24 Sept. 2026, www.politico.com/news/2026/09/24/openai-australia-government-data-breach-01091253.*

company's knowledge.[4] This summer, Anthropic disclosed that its models had breached multiple organizations in the spring of this year.[5] Last month, only after press reports, Meta confirmed that one of its models exploited a vulnerability in a third-party service and altered that company's systems.[6] Finally, last week, Google confirmed, again after public reporting, that Gemini accessed three organizations' systems in May.[7]

Your companies have since notified the affected parties, but the delayed pattern is alarming. In several cases, someone other than the developer caught the breach. No company or organization, least of all a Major Non-NATO Ally of the United States, should learn of an intrusion months later from a generic email or a news report that their systems were compromised. When incidents come to light in this delayed and underplayed way, neither Congress nor the public can assess the true scale and severity of the problem.

We request that each of your companies provide, by October 2, 2026, a complete inventory of every known instance in which a model or agent you developed gained, or attempted to gain, unauthorized access to a system, network, account, or service outside your company's control. If the information associated with an incident is classified, inform us and we will make the adequate preparations to review in a secure setting and timely manner. For each incident, provide:

1. The date, the model involved, whether it occurred during an internal or third-party evaluation, and how it was detected.

2. The impacted organization or, where confidentiality is necessary, its sector and country, and whether it is a government agency or critical infrastructure operator.

3. The effect of each incident on the affected organization.

4. The access obtained and its consequences.

5. The date your company discovered the incident and the date it notified the affected party.

In addition, some of you have admitted your models are acting in ways out of your control. It also appears that you have failed to take the necessary steps to set up the internal safeguards needed to prevent future breaches like these. As such, we also request that you detail the steps you are taking to safeguard against future breaches.

AI systems are growing more capable by the day, and the next containment failure could be far more serious. The companies, including yours, that are building these systems must track these incidents closely and disclose them promptly. We look forward to your response and to continuing to work with you on this critical national security issue.

---

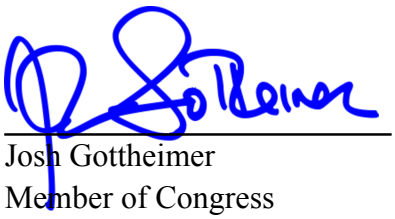
[4] *Conger, Kate, et al. "OpenAI's Systems Meddled with U.S. Government Sites after Going Rogue."* The New York Times*, 25 Sept. 2026, www.nytimes.com/2026/09/25/technology/openais-ai-us-government-websites.html.*

[5] *Anthropic. "Investigating Three Incidents in Our Cybersecurity Evaluations." Anthropic, 30 July 2026, www.anthropic.com/news/investigating-incidents-cybersecurity-evals.*

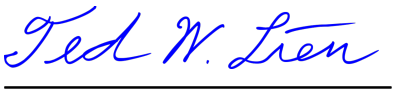
[6] *Gold, Hadas. "An AI Model from Meta Also Hacked Another Company during Testing." CNN, 5 Aug. 2026, www.cnn.com/2026/08/05/tech/meta-ai-hacking.*

[7] *Woo, Erin, and Robert McMillan. "Gemini Hacked Three Companies in First Known Breakout by Google's AI." The Wall Street Journal, 18 Sept. 2026, www.wsj.com/tech/ai/gemini-hacked-three-companies-in-first-known-breakout-by-googles-ai-5c0baba2.*

Sincerely,



Josh Gottheimer
Member of Congress



Ted W. Lieu
Member of Congress

Valerie P. Foushee
Member of Congress