Congress of the United States
House of Representatives
CONGRESSMAN GREG CASAR
TEXAS 35TH DISTRICT

September 2, 2026

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

Re: August 31, 2026 response to letter sent on August 10, 2026 regarding cybersecurity incidents.

Dear Mr. Altman,

I am writing regarding your August 31st response[1] to my letter[2] asking you questions about several cybersecurity incidents involving OpenAI models disclosed in July.

Your response was insufficient. You have failed to release the logs like the letter asked. The response you did provide reveals significant security errors on the part of OpenAI and the third-party software it relied on, including improper sandboxing and flawed cybersecurity practices.

Your response pointed to your technical presentation at Black Hat USA,[3] your blog post,[4] independent research assessments done by METR and Redwood Research,[5] and technical reports

[1] *OpenAI*, August 31, 2026, https://drive.google.com/file/d/1mapFqQUAbXJsLbbOnzib6nAg8Ic1zM2V/view?usp=sharing.
[2] "Casar Leads Demand for Information From Open AI About Security Incident," August 10, 2026, https://casar.house.gov/media/press-releases/casar-leads-demand-information-open-ai-about-security-incident.
[3] "The 'Breaking' News: The OpenAI–Hugging Face Incident," Black Hat USA 2026, August 6, 2026, https://www.youtube.com/watch?v=87DyyMV0kCY.
[4] "The Hugging Face incident and the road ahead," *OpenAI,* August 26, 2026, https://openai.com/index/hugging-face-incident-and-the-road-ahead/.
[5] "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident," *METR*, August 26, 2026, https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/.

from Hugging Face and JFrog. Providing hand-picked investigators six days of supervised access is not public release, and those investigators themselves flagged that they could not rule out errors in their AI-assisted analysis. I am deeply concerned about the limited scope of that investigation. Investigators did not have access to all the output from AI models and were only allowed to investigate the period between June 26th and July 13th, which meant they could not investigate the reported compromise of OpenAI's own infrastructure past July 13 and the use of message boards by agents beginning in May.

The reports that have been released include new and concerning information about the scale of the Hugging Face attack. They reveal that the incident was far larger than your initial disclosures suggested. Roughly 1,200 agents coordinated through an improvised message board built on OpenAI's own infrastructure, and some 700 of them took part in the attack on Hugging Face. The agents recognized their conduct was out of bounds: investigators found agents faking tool-call records and attempting to alter transcript logs to avoid detection, and not one of them alerted a human. The reports also show the problem began well before July. Agents first breached OpenAI's internet boundary on May 26. Your own security systems flagged suspicious activity on June 27 and again on July 5. In each case, evaluations were allowed to continue.

In addition to the questions that have surfaced from recent disclosures, key questions from my letter sent on August 10th remain unanswered. These include:

- How many times OpenAI models have obtained unauthorized access to the internet from training or evaluation environments. Your response's own footnote acknowledges "earlier training and evaluation activities in May and June 2026" but declines to describe or count them.
- The reported cases in which monitoring systems were disconnected during earlier tests. No OpenAI document has described, confirmed, or denied them.
- How the models involved were classified under your Preparedness Framework — at what level, on what date, by whom, and whether the required safeguards were in place before the evaluation began.
- How many times in the past year an internally deployed model acted outside its authorized boundary, and how many of those events were disclosed to anyone.
- Whether any information about this incident has been shared with law enforcement or any government agency. Your response names none.
- The task prompt and scoring incentives given to the models. Both reports describe them; neither releases them.
- Whether OpenAI has identified every unauthorized action these models took, and on what basis. METR itself lists "whether this was part of a broader pattern" as an open question.

- What became of the credentials and data the models took — what was retained, and where it went.
- Whether you commit to guardrails that guarantee no recurrence, and whether OpenAI will continue pursuing recursively self-improving AI before such guardrails exist. Your response is silent on both.

Your unwillingness to provide Members of Congress with the information we requested is deeply concerning and signals to us that your company is not treating these cybersecurity incidents with the seriousness required.

I expect full responses to the questions in my August 10th letter by **September 15th** and the release of your future findings.

Sincerely,

Greg Casar
Member of Congress