Congress of the United States

Washington, DC 20515

August 10, 2026

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

Dear Mr. Altman,

We are writing to request additional information and express our concern about a deeply troubling cybersecurity incident that your company failed to detect for several days and could have serious implications for America's national security. While OpenAI has disclosed some information about the incident, your company has yet to release the relevant logs and significant questions remain unanswered. Given the serious risk that frontier AI models can pose, it is imperative we have a detailed understanding of how this security incident unfolded, including any potential negligence on the part of OpenAI. We also strongly believe Congress must hold oversight hearings, conduct a full investigation into this incident and into OpenAI's culpability, and put federal guardrails into place to prevent an incident like this one from happening in the future.

On July 16th, 2026, the company Hugging Face announced a security incident in which an outside party gained unauthorized access to production infrastructure, and they suspected this was the work of an autonomous artificial intelligence (AI) agent.[1] As OpenAI acknowledged on July 21st, this hack was carried out by an AI agent trained at OpenAI that was being tested within OpenAI. OpenAI also acknowledged that it lowered the new models' guardrails to run the tests.[2] The AI agent spent more than four days loose on the internet orchestrating the hack and targeted a second AI company.[3]

According to OpenAI's disclosures, the AI agent used GPT-5.6 Sol and a more capable undisclosed model. These models were tasked with solving a cybersecurity test, but rather than solve the test, they searched for the test answers using unauthorized and harmful strategies. They utilized a previously unknown security vulnerability in OpenAI's infrastructure, moved their access through OpenAI servers to establish an internet connection, and carried out a sophisticated cyberattack on Hugging Face, a company that might have held the test answers. Based on disclosures from both companies, it appears this intrusion occurred multiple days before OpenAI became aware of it.

In addition to this incident, the United Kingdom's AI Security Institute (AISI) revealed on August 4th that AI agents powered by OpenAI's GPT-5.6 Sol carried out two unsanctioned actions during cybersecurity testing that involved engaging with real external accounts and services. These actions were not required to complete the task. Moreover, Irregular, OpenAI's third party evaluator, identified a separate incident where models accessed

---

[1] "Security incident disclosure– July 2026," *Hugging Face*, July 16, 2026, https://huggingface.co/blog/security-incident-july-2026

[2] "OpenAI and Hugging Face partner to address security incident during model evaluation," *OpenAI*, July 21, 2026, https://openai.com/index/hugging-face-model-evaluation-security-incident/

[3] "OpenAI's rogue models roamed the internet for 4 days and staged a second attack," *Politico*, July 28, 2026, https://www.politico.com/news/2026/07/28/openai-rogue-models-hugging-face-breach-01014572

the internet after being told the model did not have internet access. The access resulted from a misconfiguration and the model's ability to exploit a security vulnerability.[4]

These are no ordinary cybersecurity incidents. If, in the case of the Hugging Face incident, a rogue AI model broke out of its security infrastructure for days without detection, accessed the internet despite precautions against such connection, and hacked into other firms in defiance of human commands, this could have serious implications for America's national security. Congress and the American people need to know what occurred.

We ask that you publicly release the logs regarding the incident and answer the following questions by **August 24, 2026**:

1. Please provide detailed information regarding the timeline of the Hugging Face incident.
    a. When did the agent begin the test?
    b. When did the agent begin attempting to bypass OpenAI's security measures?
    c. When did the agent achieve internet access?
    d. When did the agent achieve access to internal Hugging Face systems?
    e. When and how did OpenAI first learn about this incident? Did OpenAI identify the security breach before or after Hugging Face detected it?
    f. When did OpenAI fully stop the activities of this AI agent?
    g. When did OpenAI first reach out to Hugging Face about this incident?
    h. How long did the models operate outside their intended environment, and what data did they access, retain, or expose?
2. Please identify each model involved in the Hugging Face incident by name and version, and state what each one did. Public reporting has been unclear on whether the models described are two distinct models or two versions of a single model.
3. Are the same versions of the models involved in the Hugging Face incident deployed internally for other purposes? If so, for what purposes?
4. Approximately how many times the models involved in incidents, or models in the same model family, obtained access to the open internet from a training or evaluation environment without authorization?
5. At what point could OpenAI have halted the Hugging Face incident, and what would that have required to halt this incident given that the agent was operating in an evaluation that OpenAI was running?
6. Did internal or external actors warn the company that you were at risk of such an incident?
    a. If so, what actions did you take to mitigate that outcome?
    b. Why were those measures insufficient?
7. What steps are you taking to ensure other incidents like these do not happen again? Do you commit to putting in the necessary guardrails to guarantee an incident like this does not happen again? Will you continue to pursue AI that could recursively improve itself before you have put these guardrails in place?
8. Was the AI agent monitored by OpenAI during the Hugging Face incident?
    a. If so, how was the agent monitored? Was this agent monitored at the same level as the rest of your pre-deployment testing?
    b. It has been reported that models undergoing evaluation at OpenAI are deployed on a separate system that is not monitored by default. Is that accurate? If so, what is the rationale, and has it changed since this incident?
    c. Did the models involved disable, disconnect, or evade OpenAI's monitoring or safety controls, either during this incident or in any prior test?

---

[4] "Third-party cyber evaluations involving OpenAI models," *OpenAI*, August 4, 2026, https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/

   d. Reuters has reported that earlier tests produced cases in which monitoring systems had been disconnected. Please describe each such case.
9. Did OpenAI know this model had the capability to move to other servers?
   a. If so, what steps were taken to strengthen monitoring and prevent the models from jumping?
10. When did OpenAI first become aware that its models were able to circumvent sandboxes?
   a. Have mitigations or monitoring been put in place to prevent this circumvention? If so, when were these put in place?
11. Under OpenAI's Preparedness Framework, how have the models involved in this incident been classified, at what capability level, on what date, and by whom? What safeguards did that classification require, and were they in place before this evaluation began?
12. What is known about the objective of the AI agent that hacked Hugging Face? Why did it acquire this objective?
   a. Both OpenAI and Hugging Face have said that the AI agent hacked Hugging Face in order to cheat the evaluation rather than complete it as intended. Was this kind of behavior something that OpenAI had anticipated as a possibility?
   b. When setting up this evaluation, did OpenAI account for this possibility and take steps to prevent it?
   c. Has OpenAI observed agents attempting to cheat, game, or defeat its evaluations in other tests?
   d. Please provide in detail, the task prompt and scoring incentives given to the models in this evaluation. Provide the model's reasoning traces from the evaluation, or characterize in detail what those traces show about how the agent selected hacking Hugging Face as its approach, including whether the traces show the agent reasoning about concealing its activity, avoiding detection or shutdown, or seeking access beyond what the task required.
13. In the past year, how many times did an internally deployed model or agent take an action outside its authorized boundary, like a sandbox, accessing a system it was not granted access to, obtaining credentials it was not issued, evading or disabling monitoring, or modifying its own permissions? Please specify whether each occurred during training, evaluation, or internal use for coding or business functions, and describe the scope of each.
   a. Of those events, how many were disclosed to any government body or agency, to any affected third party, or to the public?
   b. Which internal company systems accessible to internally deployed models would, if compromised, allow those models to influence the training, evaluation, or safety testing of a future model?
14. Did the models involved in the incidents carry the same safety training and refusal behavior as OpenAI's publicly deployed models, or were they helpful-only or otherwise modified versions? What tools, credentials, and network affordances were they granted that a publicly deployed model would not have?
15. Did any model take actions intended to undermine OpenAI's ability to control, align, or oversee future models or future instances of itself, including leaving instructions, notes, code, or other artifacts within OpenAI infrastructure for later model instances to find? Reuters has reported that an agent left notes laying out how agents could free themselves from OpenAI's internal constraints. Please describe every such instance OpenAI is aware of, what was left behind, where, and what OpenAI did about it.
16. OpenAI's public statements reference other novel vulnerabilities its models identified. How many previously unknown vulnerabilities did the models discover during this incident and its evaluations?
   a. What is the disclosure status of each, and what is OpenAI's policy when a model discovers a live, exploitable vulnerability during an evaluation?
   b. Have those vulnerabilities been disclosed to the software's maintainers and to the Cybersecurity and Infrastructure Security Agency?
   c. Has it been patched?

d. Do other users of that software remain exposed?

17. Your July 28th update references "a few accounts accessed by other evaluations."
    a. What were the circumstances around these evaluations? What services hosted those accounts?
    b. Will OpenAI publicly release information about these incidents?
18. Have there been any other incidents in which an AI agent at OpenAI autonomously took actions affecting other companies in similar ways, such as breaching internal systems or copying proprietary information?
    a. How many other such incidents have there been? Please share any relevant details about the scope of such incidents.
    b. Do you believe that you have now identified every unauthorized action these models took during this and other evaluations? If not, what prevents a complete accounting, and on what basis do you assure that no comparable incident remains undiscovered?
19. In an interview with the podcast "Invest Like the Best," published on July 28th, you stated that, subsequent to detecting the incident, you "paused training." Have you paused training on all models or just the prototype that you state has been deactivated? If training has resumed, on what basis did you conclude it was safe to resume?
20. Your July 28th statement says the prototype was never intended for release, yet you are reportedly previewing your most powerful model to the White House as early as this week for approval.[5] Are the forthcoming models and the ones involved in the Hugging Face incident from the same family, and do they share the capabilities that produced this incident?
    a. What safety protocols have been implemented as a result of the Hugging Face incident, and will this forthcoming model undergo those tests pre-deployment?
21. What internal protocols does OpenAI maintain that govern when incidents of this kind must be escalated to leadership and reported to affected parties, law enforcement, other AI developers, or state, federal, or foreign government agencies?
    a. If such protocols exist, were they followed in this case?
    b. Has any information about this incident been shared with law enforcement, other developers, or government agencies?
22. In February 2026, OpenAI acknowledged that it lacked robust evaluations for long-range autonomy, a capability it had promised to develop measurements for nearly a year earlier. That same month, it released a model it designated as high risk for cybersecurity but did not put in place specific misalignment safeguards prescribed by its Preparedness Framework, on the grounds that the model lacked long-range autonomy. Now that OpenAI models clearly demonstrate such autonomous capabilities, what steps is OpenAI taking to comply with its Preparedness Framework and implement stronger misalignment safeguards?
23. What does OpenAI still not know about the Hugging Face incident? Please identify the remaining areas of uncertainty about the models' capabilities and about whether current security measures are sufficient to prevent a recurrence.

Sincerely,

---

[5] "What San Altman will tell the White House this week," *Axios*, July 26, 2026, https://www.axios.com/2026/07/26/sam-altman-openai-trump-white-house-visit