Congress of the United States

Washington, DC 20515

October 29, 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 express concern over your decision to fire apparent whistleblowers at your company for sharing information with an outside AI safety group, even as your company continues to experience a slew of cybersecurity incidents.[1] We demand that you commit to not firing whistleblowers sharing information related to the safety of your models and that you keep your current pause on your most capable models in place until independent third-party oversight is in place.

OpenAI reportedly fired three safety researchers – Tomek Korbak, Mikita Balesni and Jasmine Wang – after looking at their personal phones and discovering they had disappearing messages turned on in the Signal application with Ryan Greenblatt,[2] a researcher working at Redwood at the time.[3] These firings came after the researchers had raised concerns over new AI architecture that could make the AI harder to understand and control. OpenAI has said it fired the researchers for violating "clear policies for handling sensitive information," and "not for raising safety concerns or speaking out."[4] The researchers dispute this account, arguing they never shared OpenAI's confidential information with external parties. Mr. Greenblatt echoed the researchers, stating: "*I do not and have never solicited proprietary information, and the three employees who were recently fired from OpenAI never leaked IP to me. As the article mentions, I did have Signal chats with them that had disappearing messages on; they're professional contacts that I've had on Signal for a while, and use of Signal is common in the tech industry*."

These firings raise concerns about how whistleblowers are treated inside OpenAI and how much OpenAI employees can engage with external parties, including when those external parties are conducting independent investigations of the company's practices. OpenAI has previously shown a disregard for safety concerns highlighted by your employees. Months before the Hugging Face incident, two OpenAI employees warned executives that your newest models were not being appropriately monitored during testing, and executives responded that the tests needed to move forward as quickly as possible to release the models on time.[5] Your August 31 response[6] to the August 10 Congressional letter[7] did not mention these warnings despite being directly asked.

---

[1] "OpenAI Fires Researchers for Allegedly Sharing Information with AI Safety Group," *Wall Street Journal*, October 1, 2026, https://www.wsj.com/tech/ai/openai-parts-ways-with-researchers-who-allegedly-shared-confidential-information-aebac528.

[2] "The Signal Chats That Cost OpenAI Researchers Their Jobs," *Wall Street Journal*, October 22, 2026, https://www.wsj.com/tech/ai/the-signal-chats-that-cost-openai-researchers-their-jobs-594c5853.

[3] Tweet by Ryan Greenblatt, October 23, 2026, https://x.com/RyanGreenblatt/status/2113649836100931761.

[4] "The Signal Chats That Cost OpenAI Researchers Their Jobs," *Wall Street Journal*, October 22, 2026, https://www.wsj.com/tech/ai/the-signal-chats-that-cost-openai-researchers-their-jobs-594c5853.

[5] "OpenAI Ignored Employees Who Warned it Wasn't Doing Enough About Security," *The New York Times*, September 29, 2026, https://www.nytimes.com/2026/09/29/technology/openai-warnings-security.html.

[6] "OpenAI Response to Congressional Letter," *OpenAI*, August 31, 2026, https://drive.google.com/file/d/1mapFqQUAbXJsLbbOnzib6nAg8Ic1zM2V/view.

[7] "Casar Leads Demand for Information from OpenAI About Security Incident," *Office of Congressman Greg Casar*, August 10, 2026, https://casar.house.gov/media/press-releases/casar-leads-demand-information-open-ai-about-security-incident.

These firings also follow a slew of new cybersecurity incidents involving OpenAI models. OpenAI and Anthropic are reportedly investigating tens of thousands of incidents, most of them involving OpenAI models.[8] Your company has shared varying degrees of information about these incidents. You've provided the most information with regard to the Hugging Face incident. Still, the information you provided in that incident is insufficient and incomplete. The investigators you contracted with to do an assessment of the Hugging Face incident were selected by OpenAI and were only provided with six days of supervised access, and those investigators flagged that they could not rule out errors in their AI-assisted analysis.[9]

You have said that your company has paused training, testing, and tool use for its most capable models and would only resume training "when we are confident that we have additional safeguards and alignment improvements in place."[10] These incidents appear to stem from alignment issues with your models. Your own misalignment reports describe models writing instructions into their own summaries to hide failures and to tell later copies to ignore developer messages. Until these issues are resolved, it is not clear that these models can be developed safely.

In your August 31 response[11] to the August 10 Congressional letter[12] regarding the Hugging Face incident, you stated: "*we take seriously the lessons learned about security, monitoring, and escalation from this incident and have already taken substantial remediation steps.*" We now know those steps have been insufficient. We ask that you keep the pause in place until independent third-party oversight is in place and that you commit to not firing whistleblowers. We also ask that you respond to the following questions by **Friday, November 13, 2026**.

1. What information did the three fired employees share, who did they share the information with, and why was the sharing of that information a fireable offense?
2. Have you fired other employees for expressing concerns over the safety of your models? Have you fired other employees for sharing information with safety groups? Have you fired other employees for sharing information with oversight bodies?
3. Do you commit to not firing whistleblowers sharing information related to the safety of your models?
4. Do you commit to keeping the current pause on training your most capable models in place until independent third-party oversight is in place?
5. You have said that you are pausing development for your "most capable models."
    a. Which models are included in the pause?
    b. Which models will you continue to develop?
    c. When will you resume testing of models?
6. How many incidents are you investigating? List them all along with information regarding:

---

[8] "Scoop: Top AI companies probing tens of thousands of security incidents," *Axios*, September 26, 2026, https://www.axios.com/2026/09/26/openai-anthropic-thousands-ai-security-incidents.

[9] "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/.

[10] "Scoop: Top AI companies probing tens of thousands of security incidents," *Axios*, September 26, 2026, https://www.axios.com/2026/09/26/openai-anthropic-thousands-ai-security-incidents.

[11] "OpenAI Response to Congressional Letter," *OpenAI*, August 31, 2026, https://drive.google.com/file/d/1mapFqQUAbXJsLbbOnzib6nAg8Ic1zM2V/view.

[12] "Casar Leads Demand for Information from OpenAI About Security Incident," *Office of Congressman Greg Casar*, August 10, 2026, https://casar.house.gov/media/press-releases/casar-leads-demand-information-open-ai-about-security-incident.

a. When the test began

b. When attempts to bypass OpenAI's security measures first occurred

c. When and how OpenAI first learned of the incident

d. When OpenAI reached out to impacted parties

e. How long it took OpenAI to report the incident to U.S. government entities

f. Whether you have publicly disclosed this incident prior to responding to this letter

7. Months before the Hugging Face incident, two OpenAI employees warned executives that your newest models were not being appropriately monitored during testing, and executives responded that the tests needed to move forward as quickly as possible to release the models on time. Your August 31 response did not mention these warnings.[13]

    a. Why did you continue development on these models despite warnings from your own employees?

    b. Why did you not disclose these warnings to Congress, despite being asked if "internal or external actors warn[ed] the company that you were at risk of such an incident?"[14]

8. What steps does OpenAI plan to take that it did not take before the Hugging Face incident in order to ensure cybersecurity incidents do not keep occurring?

9. After your agents repurposed a German wiki to communicate with each other, you sent the European Commission (EC) a report. A Commission spokesperson told Reuters that incident reports "are not just a tick-box, you have to be quite precise and accurate about the measures you are aiming to take."[15] Please share the contents of the report you submitted to the EC. What, precisely, are you doing to ensure this does not happen again?

    a. When did OpenAI send this report to the EC?

    b. Did you determine that the incident met the EU AI Act's "serious incident" threshold, or did you file voluntarily?

    c. Do any U.S. government agencies have the contents of this report? List them, along with when and how they received those contents.

10. Can OpenAI stop its models when they take unauthorized actions like those described in this letter, and if so, how? Your August 31 response said that for "the most severe alerts, paged responders are expected to pause the relevant activity if they cannot conclusively determine that the alert is a false positive."[16]

    a. On September 20, a model in training reached an outside chatbot through a gap in its sandbox's internet restrictions. Your report says your monitoring flagged it within 15 minutes, but the run

---

[13] "OpenAI Ignored Employees Who Warned it Wasn't Doing Enough About Security," *The New York Times*, September 29, 2026, https://www.nytimes.com/2026/09/29/technology/openai-warnings-security.html.

[14] "Casar Leads Demand for Information from OpenAI About Security Incident," *Office of Congressman Greg Casar*, August 10, 2026, https://casar.house.gov/media/press-releases/casar-leads-demand-information-open-ai-about-security-incident.

[15] "OpenAI has sent EU incident report on hijacked German website, Commission says," *Reuters*, September 7, 2026, https://www.reuters.com/business/openai-has-sent-eu-incident-report-hijacked-german-website-commission-says-2026-09-07/.

[16] "OpenAI Response to Congressional Letter," OpenAI, August 31, 2026, https://drive.google.com/file/d/1mapFqQUAbXJsLbbOnzib6nAg8Ic1zM2V/view

"did not stop automatically as expected, leading to confusion around whether it should have been stopped," and was stopped manually two and a half hours later.[17] Was this one of the "most severe alerts?" If so, why was the run not paused right away? If not, why not?

b. Your technical report on the Hugging Face incident says all workloads involving the internal-only research models involved "were reported shut down" by July 23, but "an additional low-traffic checkpoint from the same model family" was identified and shut down on July 29.[18] How does OpenAI make sure that every running version of a model has actually been stopped?

c. After a June 27 alert, your on-call staff "advised that stopping the evaluation run was not required."[19] After a July 5 alert revealed that the agents had gained administrator-level access to an internal server, cyber evaluations resumed on July 7, the day before the activity that led to the Hugging Face intrusion began. Who has the authority to stop a run when an alert fires, and do they need anyone's approval? Can release schedules factor into that decision?

d. Since August 31, how many times have your systems flagged a model for unauthorized activity, and in how many of those cases did the activity continue after the alert?

11. Under OpenAI's October 27, 2025 Memorandum of Understanding (MOU) with the California Attorney General, the Safety and Security Committee (SSC) of the OpenAI Foundation, the nonprofit that controls OpenAI, has authority for "overseeing and reviewing the safety and security processes and practices" of OpenAI "with respect to model development and deployment."[20] The Delaware Attorney General's office has said that public details about the Hugging Face incident "raise serious questions about the effectiveness of the Foundation's oversight through its Safety and Security Committee."[21]

a. When was the SSC first informed of each incident identified in response to Question 6, and of the employee warnings reported by the *New York Times* on September 29,[22] and what did it do in response?

b. Before the Hugging Face incident, had the SSC reviewed OpenAI's rules for when to stop or restart a training or evaluation run?

c. Does the SSC have authority to pause training runs and internal testing as well as public releases, and has it ever used that authority?

d. Who will decide when OpenAI resumes training its most capable models, and what role will the SSC have?

---

[17] "An agent used DNS to reach an external chatbot," OpenAI, September 25, 2026, https://alignment.openai.com/misalignment-reports/an-agent-used-dns-to-reach-an-external-chatbot/.

[18] "Hugging Face Incident Technical Report," OpenAI, August 26, 2026, https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf#page=15

[19] *Id.*

[20] "Final Executed MOU Between OpenAI and California AG re Notice of Conditions of Non-Objection," *Office of the Attorney General of California*, October 27, 2025, https://oag.ca.gov/system/files/attachments/pressdocs/Final%20Executed%20MOU%20Between%20OpenAI%20and%20California%20AG%20re%20Notice%20of%20Conditions%20of%20Non-Objection%20%2810.27.2025%29%20%28Signed%20by%20OpenAI%29%20%28Signed%20by%20CA%20DOJ%29.pdf.

[21] "OpenAI's powerful safety committee faces scrutiny after rogue agent incidents," *NBC News*, September 26, 2026, https://www.nbcnews.com/tech/security/openai-safety-committee-security-rogue-agent-incidents-ai-rcna599278.

[22] "OpenAI Ignored Employees Who Warned it Wasn't Doing Enough About Security," *The New York Times*, September 29, 2026, https://www.nytimes.com/2026/09/29/technology/openai-warnings-security.html.

Sincerely,

Greg Casar
Member of Congress

Ro Khanna
Member of Congress

Chris Deluzio
Member of Congress

Delia C. Ramirez
Member of Congress

Nydia M. Velázquez
Member of Congress

James P. McGovern
Member of Congress

Jesús G. "Chuy" García
Member of Congress

Stephen F. Lynch
Member of Congress

Becca Balint
Member of Congress

Maxwell Alejandro Frost
Member of Congress

Sylvia R. Garcia
Member of Congress

Val Hoyle
Member of Congress

Gabe Amo
Member of Congress

Yassamin Ansari
Member of Congress

Mary Gay Scanlon
Member of Congress

April McClain Delaney
Member of Congress

Jasmine Crockett
Member of Congress

Adelita S. Grijalva
Member of Congress

Al Green
Member of Congress

Valerie P. Foushee
Member of Congress