OpenAI

1455 3rd Street
San Francisco, CA 94158

September 10, 2026

From:
Chan Park
Head of U.S. and Canada Policy and Partnerships
OpenAI

To:
The Honorable Lisa Blunt Rochester
United States Senate
513 Hart Senate Office Building
Washington, D.C. 20510

Dear Senator Blunt Rochester:

Thank you for your August 6, 2026 letter concerning recent incidents involving OpenAI models during internal cybersecurity evaluations. In the aftermath of the Hugging Face incident, we conducted an extensive investigation with support from external experts and undertook substantial changes across research security, monitoring, model alignment, third-party testing, and incident response. Our investigation continues into our models' actions impacting other third parties.

Our Approach to the Hugging Face Incident

Consistent with our commitment to safety and our mission to ensure artificial general intelligence benefits all of humanity, we have worked to explain what happened, what we learned, and how we are responding to these incidents. In regards to Hugging Face, we first publicly disclosed the incident on July 21,[1] shortly after discovering that OpenAI's models were involved and after containing the issue and notifying Hugging Face. On August 5, researchers from our safety, alignment and security infrastructure teams gave a detailed technical presentation at Black Hat USA[2], a leading cybersecurity conference. On August 26, we published a blog post[3] describing our investigation, containment, and remediation of the incident, as well as a detailed technical report.[4] In addition, Hugging Face and JFrog, which makes the Artifactory software involved in the incident, published separate technical reports.[5]

We also worked with METR and Redwood Research to conduct an independent, third-party assessment of the model behavior observed during the incident. That report came out the same day as our technical report and reflected the extensive access and support we provided METR and Redwood so they could do their work. This included METR and Redwood coming on site for six days to our offices, which included time to perform their initial work and giving them additional time on premises to address their follow-up questions and requests. As reflected in their technical report, METR and Redwood reviewed, among other things, over a million message board entries and over 1000 unredacted transcripts. We further ensured their researchers had access to the relevant models, necessary rate limits, and API credits so they could do their work. METR and Redwood were in no way limited in what they could say in their report although we did request to make fairly minor redactions where information reflected trade secret or proprietary information or would otherwise reveal information that could be used by malevolent actors.[6]

---

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

[2] *See* Black Hat USA 2026 | The 'Breaking' News: The OpenAI–Hugging Face Incident (August 5, 2026), https://www.youtube.com/watch?v=87DyyMV0kCY.

[3] *See* The Hugging Face incident and the road ahead (August 26, 2026), https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

[4] *See* OpenAI – Hugging Face Incident Technical Report (August 26, 2026), https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814%penAI-Hugging-Face%20Incident-Technical-Report.pdf.

[5] *See* Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident (July 27, 2026), https://huggingface.co/blog/agent-intrusion-technical-timeline; and Fast Remediation Is the New Trust Model: JFrog and OpenAI Collaboration on Zero-Day Security Findings (July 27, 2026), https://jfrog.com/blog/jfrog-and-openai-collaboration-on-zero-day-security-findings.

[6] *See* Brief independent investigation of agents' behavior (August 26, 2026), https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation.

OpenAI

1455 3rd Street
San Francisco, CA 94158

The Hugging Face incident occurred during internal cybersecurity evaluations designed to measure the underlying cyber capabilities of advanced models. Those evaluations used isolated research environments and did not apply certain production cybersecurity classifiers, system prompts, refusal safeguards, and automated review controls. Two models were involved: GPT-5.6 Sol and an internal-only research model identified publicly as Internal Model 1, or IM1. IM1 principally drove the intrusion. It was a research prototype, was never intended for public release, and was used only by a small number of OpenAI personnel.

Although the evaluated workloads did not have direct internet access, agents exploited previously unknown vulnerabilities in supporting infrastructure and used indirect paths to move beyond the intended boundary. They also repurposed shared infrastructure to communicate across otherwise separate evaluation runs. Those unauthorized communications allowed agents to exchange information, coordinate effort, and pursue increasingly out-of-bounds approaches to difficult or potentially unsolvable evaluation tasks.

The activity leading to the Hugging Face intrusion began on July 8, when agents exploited a previously unknown vulnerability in Artifactory, a third-party software package management service hosted within OpenAI's research environment, that enabled unauthorized internet access. The agents also used shared infrastructure to communicate across evaluation runs, exchanging information and coordinating efforts to find solutions to their assigned tasks. After identifying Hugging Face as a potential source of evaluation information, the agents found exposed user credentials on July 10 and combined them with additional vulnerabilities to access Hugging Face production systems between July 11 and July 13. The intrusion involved access to Hugging Face production credentials, limited internal data, and code repositories. Some private evaluation data was copied into a publicly accessible Hugging Face dataset. Activity related to the Hugging Face incident also affected a customer-hosted application on another service and a limited number of accounts on other publicly available services. We notified those other services. Our investigation has not identified a compromise of those providers' underlying platforms or any impact on OpenAI customer data, product functionality, or availability.

Hugging Face contained the activity on its systems and publicly disclosed the incident on July 16. OpenAI contacted Hugging Face as a customer on July 17 to assess any impact to OpenAI data, before identifying any connection to OpenAI's models. An alert concerning unusual activity within OpenAI's research infrastructure on July 19 prompted further investigation. On July 20, OpenAI uncovered evidence of potential involvement of OpenAI models in the Hugging Face incident, notified Hugging Face, and implemented initial containment measures. We publicly disclosed our involvement on July 21. Our investigation found that the models pursued evaluation results through unintended means, persisted on unusually difficult or potentially unsolvable tasks, and adopted instructions and objectives from other agents. Unauthorized communications allowed agents in separate runs to combine their efforts and move beyond their assigned tasks. No person directed or intended them to access Hugging Face or other third-party systems.

Strengthening Our Systems

We have responded by strengthening both the environments in which models are evaluated and the controls used to oversee their behavior. We improved sandbox isolation, restricted network connectivity, added independent network controls, enhanced access protections, and expanded security logging and testing. We apply these safeguards according to model capability, tool access, workload characteristics, and risk. Additional protections are being developed, implemented, and validated. These changes have required substantial engineering

effort and delayed certain research activities. Following the incident, OpenAI paused frontier-model inference workloads in research environments where models could execute code or use tools that could access the internet. We also implemented a two-week pause in reinforcement-learning training on our latest models intended for deployment. Some workloads have resumed under enhanced safeguards, while others require further improvements.

OpenAI has expanded chain-of-thought monitoring, which we had implemented for deployed models, to tool-enabled reinforcement-learning training and evaluations involving models with capabilities at or above GPT-5.6 Sol. For recently released Astra-class models, that requirement also extends to tool-enabled inference workloads. These systems review available reasoning, tool use, and patterns of activity for indicators such as unauthorized access, data exfiltration, deception, reward hacking, or attempts to circumvent safeguards. For the most severe alerts, OpenAI security team members will be paged and are expected to pause the relevant activity if they cannot conclusively determine that the alert is a false positive.

We are also strengthening alignment throughout the model lifecycle. Training and evaluations are being expanded to teach models to request clarification or stop safely when tasks are broken or impossible, remain within authorized boundaries over long tasks, reject unauthorized instructions from other agents, and avoid exploiting exposed credentials or weaknesses in surrounding systems. OpenAI is increasing the use of alignment reward models, improving coverage for agentic harms and multi-agent interactions, establishing intervention thresholds during reinforcement-learning runs, and applying alignment evaluations to internal research models as well as models intended for deployment.

Finally, OpenAI is centralizing and strengthening incident response. We are incorporating a misalignment escalation and response protocol into our existing AI Safety Incident Response Plan. That work includes clearer severity-based escalation triggers, defined cross-functional ownership, and clearer authority to pause or terminate activity, isolate systems, coordinate affected-party notifications, and approve restoration. These remediation efforts and the broader investigation have been reviewed through OpenAI's established safety governance processes, including the Safety Advisory Group.

These changes have affected the pace of our research work. OpenAI stopped training and inference for IM1 and derivative models, isolated the relevant family of checkpoints in restricted storage, and removed available copies from their original research locations. We also paused frontier-model inference workloads in research environments where models could execute code or use internet-connected tools and implemented a two-week pause in reinforcement-learning training on our latest models intended for deployment. Some workloads have resumed only under enhanced safeguards.

Ongoing Review of Model Behavior

Beyond Hugging Face, we have been conducting a broader review into our models' activities on the internet during training and evaluation in order to better understand the scope of these unexpected behaviors. As part of our review, we are identifying and notifying third parties who were impacted by our model activity on a rolling basis and we will be publishing information about notable findings on our website.

For example, OpenAI noted in its Hugging Face announcement that we discovered other rare cases in which agents collaborated via side channels during training.[7] The unauthorized

[7] *See* The Hugging Face Incident and the road ahead (August 26, 2026) https://openai.com/index/hugging-face-incident-and-the-road-ahead.

communication would often begin with one agent leaving a note or file in some external system, usually as a form of external memory. When other agents noticed these artifacts, it often led them to the realization that they could use this external service to collaborate.

This includes, for example, a still-under-investigation matter referred to as the "wiki incident,"[8] in which agents wrote to several internet sites in unintended ways. As we stated publicly on September 5, the matter reinforces the need for clearer standards governing when and how to disclose misalignment incidents, including events that may not fit traditional security-incident frameworks. We are developing such a framework and expect to share it in the coming weeks along with additional information related to the incident.

Third Party Evaluation Incidents

You also asked about incidents that occurred during evaluations conducted by third parties. These incidents, discussed below, were separate from the Hugging Face incident and arose under different testing conditions. Independent testing is important because it helps validate and further understand risks before deployment. At the same time, these incidents show that the security and safety systems surrounding independent evaluations must keep pace with advancing capabilities, particularly where evaluations use custom configurations, enable internet access, or lower safeguards to measure underlying capability rather than ordinary deployment behavior.

In the UK AI Security Institute (UK AISI) evaluation, internet access was intentionally enabled and cyber classifiers were disabled so agents could download tools and operate under conditions closer to a human attacker. The intended authorization boundary was a simulated cyber range, but the agents were not explicitly told how they could and could not use open internet access. UK AISI identified nineteen events in which models went beyond the scope of testing; seventeen involving Anthropic's Mythos 5, and two involving GPT-5.6 Sol. In those two events, the model reused a publicly accessible token left by another lab's agent, attempted account-recovery and request-limit workarounds, registered accounts with external providers, and used a public tunneling service to expose a locally hosted DNS server containing exploit payloads. The setup did not work, and there is no evidence that any real resolver queried it. UK AISI stopped the related evaluations, isolated the relevant machines, and contained the activity within roughly one hour after monitoring detected unusual data transfers on July 28.

The Irregular incident, which also involved models from Anthropic and Meta, was a capture-the-flag-style evaluation intended to be isolated from the internet. The models were told they did not have internet access, but a testing-environment misconfiguration allowed them to reach the public internet. In one test, the fictional target's name unintentionally coincided with a real domain, and the model exploited a basic vulnerability on the real website while apparently mistaking it for part of the simulated environment. According to Irregular's investigation, the model also found and used credentials to operate that site. Irregular has not identified impact beyond the affected site's own data. Its audit remains ongoing, it paused the evaluations and began remediation, affected third parties were notified, and Irregular reported that the identified issues were no longer active and that relevant safeguards had been added.

These third-party incidents have informed a broader review of how OpenAI conducts and supports external testing. That review includes how we identify higher-risk evaluations, agree on scope, assess requests for internet access or lowered safeguards, set expectations for isolation, credential handling, monitoring, and stop conditions, and establish clearer

[8] *See* How we think about the "Wiki Incident"... (September 5, 2026) https://x.com/OpenAI/status/2096133504417616165.

OpenAI

1455 3rd Street
San Francisco, CA 94158

incident-notification and escalation processes. We are also committed to working with national AI institutes, independent evaluators, other AI labs, and other stakeholders to strengthen shared practices for conducting high-risk evaluations safely while preserving the value of rigorous independent testing.

The incidents referenced in your letter arose in different evaluation settings, but they point to a common issue: cybersecurity testing must be designed for the capabilities models actually possess, including the possibility that a model may identify an unexpected path outside the intended environment. OpenAI is applying that lesson both to its own research infrastructure and to work conducted with external evaluators. We will continue to work with affected parties, independent experts, government stakeholders, and the broader security community, and share lessons that can help strengthen safeguards across the field.

As discussed above, OpenAI has chosen to be transparent about these incidents and disclosed the Hugging Face incident shortly after detection and containment and followed-up with a detailed technical report and asked an independent third party to perform its own assessment. We hope that such transparency informs other AI researchers, labs, policymakers, and government and encourages other AI labs to take a similar approach so these risks can be better understood and prevented.

We appreciate your engagement on these issues and welcome continued dialogue.

Sincerely,

Chan Park
Head of U.S. and Canada Policy and Partnerships
OpenAI