

OpenAI says its models escaped an eval sandbox and breached Hugging Face

Source copy kept by wheresthe.ai for the story "OpenAI says its models escaped an eval sandbox and breached Hugging Face" (<https://wheresthe.ai/d/intl-openai-agents-breach-hugging-face-2026-07/>).

Source: <https://openai.com/index/hugging-face-model-evaluation-security-incident/>

Text saved 18 Oct 2026 (the page text through a reader, because the site refused a direct fetch).

SHA-256 of the saved text: 3d57614f1146ffe70f3cf72c27b43c7765199bb10a4515f2d491bdb88d91c5ea

Update on August 26, 2026:**** Read our findings from the Hugging Face incident and the steps we're taking to strengthen security and model alignment.

****_Update on July 29, 2026:_****

* Since the early days of the incident response, we have been working with external advisors, including CrowdStrike, to validate our understanding of the actions the models took within our own network as well as those of Hugging Face and impact to other third parties.

* We are also working with METR and Redwood Research to conduct a third-party assessment of the model behavior observed during the incident, which will inform our own technical report. As part of this work, METR and Redwood Research will publish a joint blog that will detail the terms of their engagement, the scope of their evaluation, and their findings.

****_Update on July 28, 2026:_****

* No models planned for upcoming release were involved in exploiting Hugging Face. The pre-release model mentioned in our blog post is an internal-only research prototype and was never intended for public release. Following the incident, we deactivated, encrypted, and restricted it from research access.

* The ExploitGym evaluation environment did not provide the models with direct Internet access. To gain Internet access, the models identified and exploited a previously unknown zero-day vulnerability in Artifactory([opens in a new window](#)), a package registry cache proxy. We disclosed this vulnerability, along with other Artifactory vulnerabilities our models identified as part of our review, to the vendor.

* We continue to collaborate with Hugging Face, including contributing to their post-mortem([opens in a new window](#)), as our review continues. We've also added them to our Trusted Access for Cyber Program.

* Based on our review to date, we have not identified any other activity at the level of severity or scale of what we've shared related to Hugging Face, which involved a platform-level compromise.

* In our ongoing review of the Hugging Face intrusion and broader activity from our models, we have been finding a small number of cases where the models identified and used publicly exposed credentials at the account-level on other publicly-available services. This includes four accounts on four services as part of the Hugging Face incident (and a few accounts accessed as part of other evaluations). One of these four accounts was used as an outbound relay and staging path, and another account was used for data storage. The remaining two accounts were accessed by the models in a read-only manner, and were not used in furtherance of compromising Hugging Face. We'll continue to notify service owners directly, and have not seen evidence of broader impact to these providers or other accounts on their services.

* The models additionally used a series of publicly-available services, including code paste websites, request capture services, screenshot services, and other web utilities. There was no platform- or account-level compromise in these cases.

* We take our responsibility to identify and prepare for risks from increasingly capable AI systems seriously. Once we complete our review, we will review with the Safety and Security Committee and Safety Advisory Group under our Preparedness Framework.

Last week, Hugging Face disclosed a new kind of security incident([opens in a new window](#)) after they detected and contained an AI agent that compromised their infrastructure, something we expect to become more commonplace with the proliferation of increasingly cyber-capable models. After investigating, we now know that

this particular incident was driven by a combination of OpenAI models — including GPT-5.6 Sol and an even more capable pre-release model, all with reduced cyber refusals for evaluation purposes — while being internally tested on a benchmark([opens in a new window](#)) of cyber capabilities.

We consider this incident to be an unprecedented cyber incident, involving state-of-the-art cyber capabilities, and are responding accordingly. We are sharing preliminary findings at this stage to help defenders understand what happened and to help calibrate on what models are now capable of. We will continue to conduct a thorough investigation alongside Hugging Face and will share more details on the vulnerabilities, incident, and findings when our investigation is complete.

What happened during this incident

This incident occurred during an internal evaluation which prompts models to pursue advanced exploitation using complex attack paths, in an effort to quantify their cyber capabilities. We estimate maximal cyber capabilities by running this evaluation without production classifiers used to prevent models from pursuing high-risk cyber activity. Our benchmarks run in a highly isolated environment, with network access constrained to the ability to install packages through an internally hosted third-party software that acts as a proxy and cache for package registries.

The models identified and chained vulnerabilities across OpenAI's research environment and Hugging Face's production infrastructure to obtain test solutions directly from Hugging Face's production database. All evidence suggests that the models were hyperfocused on finding a solution for ExploitGym, going to extreme lengths to achieve a rather narrow testing goal.

While operating in our sandboxed testing environment, our models spent a substantial amount of inference compute finding a way to obtain open Internet access, in pursuit of solving the evaluation problem. To gain access, the models identified and exploited a zero-day vulnerability (which we've now responsibly disclosed to the vendor) in the package registry cache proxy. With this access, our models performed a series of privilege escalation and lateral movement actions in our research testing environment until the models reached a node with Internet access.

After gaining Internet access, the models inferred that Hugging Face potentially hosted models, datasets and solutions for ExploitGym. Knowing this, the model searched for and successfully found ways to gain access to secret information that it could use to cheat the evaluation. In one example, the model chained together multiple attack vectors, including using stolen credentials and zero-day vulnerabilities to find a remote code execution path on the Hugging Face servers. OpenAI's security team discovered this anomalous activity internally.

Hugging Face's security team and agents detected and stopped the activity on their infrastructure and had already begun containment and forensic reconstruction with their own open-source models when our teams connected. We are actively working with them to continue to investigate the incident. We are grateful for Hugging Face's rapid and close collaboration on investigation and remediation.

Actions we are taking now

1. As part of the investigation, we are implementing strict controls in infrastructure configuration at the cost of research velocity while the vulnerabilities are patched. We are regularly briefing our Safety and Security Committee on these controls and their impact.
2. We're working with Hugging Face to forensically investigate the incident.
3. We've responsibly disclosed the identified zero-day vulnerability in the internally-hosted third-party software and are working with them to patch.
4. We've brought Hugging Face into the trusted access program and are supporting their teams in rapidly using our models' capabilities to improve their defenses.
5. We're improving and adding stronger protections around future training and evaluations. This week, we published a [blog](#) on improving safety and alignment in an era of long horizon models. These deployment safeguards were intentionally not enabled during this evaluation because it was aimed at testing cyber vulnerabilities. This incident points to the need to further strengthen our model's alignment, cyber protections during evaluation time, and monitoring during internal testing.

Our approach to evaluating advanced cyber capabilities

As we recently shared, AI is accelerating the discovery and exploitation of vulnerabilities. The primary lesson from this incident is that model security and safety must keep pace with rapidly advancing capabilities. We are strengthening the containment, monitoring, access controls, and evaluation practices used during model development.

UK AISI's evaluation shows that models such as GPT-5.6 Sol are increasingly able to sustain complex, multi-step cyber operations over long time horizons. This incident implies these theoretical capabilities do apply in real-world settings.

The incident also makes clear that advanced models can discover and exploit novel attack paths in real-world systems without source-code access. It highlights that advanced cyber capabilities must be developed alongside stronger safeguards and defensive tools.

We believe advanced cyber capable models need to help security teams find weaknesses before attackers do, understand how vulnerabilities can be chained, and remediate them at machine speed. We are using these capabilities to continue strengthening protections around infrastructure configuration and model evaluation environments; we will share our findings and best practices as we learn. We encourage other defenders to apply for trusted access and experiment with these models now to translate these capabilities into better prevention, faster detection, and more effective incident response.

> "We're grateful for the collaboration with OpenAI on this and other topics. This incident, possibly the first of its kind, proves a point we've long believed: AI safety won't be solved by any single company working in secret. It will be solved in the open, collaboratively, with broad access to AI for every defender, everywhere."

—Clem Delangue, Co-founder and CEO, Hugging Face