

OpenAI pauses frontier RL training over cyber risk after Hugging Face breach

Source copy kept by wheresthe.ai for the story "OpenAI pauses frontier RL training over cyber risk after Hugging Face breach" (<https://wheresthe.ai/d/openai-pauses-frontier-rl-training-2026-08/>).

Source: <https://openai.com/index/pacing-model-development-cyber-capabilities/>

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

SHA-256 of the saved text: a291327b2ee0ed4957da98da75db2a4f6efca3a9c4dfbcad8bae013443c09870

Over the past several weeks, two developments have underscored the growing risks associated with increasingly capable AI systems: the OpenAI-Hugging Face incident and, separately, preliminary evidence that one of our upcoming models, Astra, may meet the Critical cybersecurity capability threshold under our Preparedness Framework. Together, these developments, combined with rapid progress in our internal research, have added urgency to our work on strengthening our monitoring, alignment, and containment safeguards across all stages of the training process.

As models become more capable, the risks associated with developing and testing them internally also grow. Our standards for monitoring, alignment, and security must stay ahead of those risks. We wanted to take the time necessary to meet those standards, so we temporarily slowed the pace of scaling. This included a two-week pause in reinforcement learning (RL) training on our latest models intended for deployment while we further hardened and red-teamed our research environments and expanded the coverage of our monitoring systems. Our largest planned frontier RL run remains on hold while we conduct smaller-scale training and evaluations to assess model behavior, validate our safeguards, and establish more evidence of alignment before proceeding.

Alignment—the work of making AI systems behave as intended and responsive to human oversight—has long been at the core of our research program. We now require stronger evidence of aligned behavior throughout all of training, building on research and evaluations already underway. Keeping increasingly capable systems aligned is a challenge the whole field will need to address. The signals we are seeing from upcoming model progress make clear that we need a broader approach—one that builds on and extends beyond the current Preparedness Framework.

We think it is important to be transparent about how our approach is changing. Below, we describe the changes we have already made to our research processes and infrastructure, and the work still underway.

Strengthening safeguards for more capable models

Our approach to developing more capable models rests on three reinforcing safeguards:

1. **Monitoring**, which detects and allows us to respond to concerning behavior.
2. **Alignment**, which reduces the likelihood of harmful or unauthorized actions.
3. **Security measures**, which limit what AI systems can access or affect.

We expect models to soon drive most security work, including defending against other models. This will allow all three safeguards to scale with model capability, which we see as crucial.

We apply these safeguards across research and deployment, adapting them to each model's capabilities, operating environment, and level of risk.

Securing our research environments

As frontier models gain stronger cybersecurity capabilities, we are raising the security standards for the environments in which we train and evaluate them. Meeting these standards has required substantial engineering work and has incurred great cost and delays to frontier research. Immediately following the OpenAI-Hugging Face incident*, we paused frontier model inference in research clusters for runs that could execute code or use tools that could access the internet. We restored a more limited and secure path for code execution fairly quickly, then worked through each workload individually to determine whether it could safely resume. Some research workloads resumed under these new controls, while others required additional changes.

Over the subsequent weeks, we have defined and started implementing stronger security requirements for frontier research workloads. For example:

* **Workload isolation:** We now require stronger isolation (“sandboxes”) for workloads that execute model-generated or otherwise untrusted code. This also applies to software that could be compromised while processing model outputs.

* **Network isolation:** We have implemented more controls to isolate higher-risk and untrusted workloads from the internet. We have designed these controls so that a single compromise of a workload or supporting service does not, by itself, allow for unauthorized access to the internet or other internal networks.

* **Continuous security testing:** We have reconfigured our environment to remove potentially vulnerable shared services, reduce standing privileges, and improve security and trust boundaries. We are also improving our ability to collect and monitor security logs. Finally, we are investing in automation using our models to test these boundaries continuously against simulated attacks.

Together, these controls provide defense in depth and help protect our research environment against attacks. Today, we require the strictest level of security safeguards for workloads involving Astra or cyber models, since we have determined that Astra models may have a critical level of cyber capability. These safeguards also apply to all other cyber-related workloads.

While some Astra training and evaluations meet those requirements, a significant number of workloads remain paused until they are fully migrated and enhanced to meet the new security bar. We are prioritizing safety and alignment workloads for migration to these new environments first.

Expanding chain-of-thought monitoring

Monitoring has long been an important part of our model development process, including monitoring for internal coding agents and long-running sessions. Our monitoring efforts were especially focused on internal deployments of frontier models and frontier RL training runs, which we considered the highest-risk workloads because they can combine advanced model capabilities with access to tools, sensitive systems, and internal or external networks. They were not applied to all inference and training runs using tools. We have revised and expanded our approach with our new monitoring setup.

This new monitoring setup is multistage, starting with activation classifiers (detectors that inspect a model's internal activity) that run at every sampled token and escalate potential concerns to increasingly sophisticated and high-compute automated investigators. Those systems examine tool actions, available reasoning, and the full sequence of activity for unauthorized access, data theft, destructive behavior, and attempts to defeat safeguards. We aim to issue an alert within 30 minutes after concerning activity is surfaced through our monitoring system. If the monitoring system identifies a likely violation of a critical security boundary, it generates a highest-priority alert. In our current implementation, the safety, security, and research teams are paged. If they cannot conclusively determine within 30 minutes that the flag is a false positive, those teams are expected to pause the activity.

This monitoring is required for all RL training and evaluations involving tools for models of Sol capability or higher. Once we determined that Astra may have critical cyber capabilities on August 7, we added an additional monitoring requirement for all inference of Astra with tools (not just RL training and evaluations).

These safeguards require meaningful compute. Our current estimates put monitoring overhead at roughly 20% of the inference compute being monitored, though the cost varies substantially across training and evaluation workloads. We'll share more details about this system in a forthcoming blog post.

Advancing alignment research

Alignment research is central to our mission. Our mission is to ensure that artificial general intelligence benefits all of humanity. As models gain advanced capabilities, such as the ability to cyberattack, and operate in more complex environments, misaligned behaviors like reward hacking (finding ways to receive high rewards in training without actually achieving the intended outcome), deception, or unauthorized access will create increasingly serious risk.

For RL runs on the most capable models, we are now applying our core alignment techniques across more stages of the training process. This includes improving reward models to better detect and discourage unsafe behavior across tasks and environments; training models to be more honest about their actions, capabilities, and

limitations; and reducing behaviors that exploit weaknesses in rewards, graders, tools, or oversight. We are also increasing training coverage for behaviors that could cause harm when models interact with external systems or resources.

We are continuing to invest aggressively in alignment research, increase evaluation coverage, and use what we learn to inform training and safeguards. We plan to share substantially more about our alignment research in the near future, including what we are learning about model behavior and any novel challenges we uncover.

What's next

We will evolve our Preparedness Framework to bring these safeguards together across training and deployment, and to better reflect the capabilities of future models and the environments in which they operate. Developing methods that can scale with those capabilities will require sustained investment in model-assisted security, more effective monitoring, and continued advances in alignment research. We intend to involve external organizations and share more of what we learn as our approach develops.

The capabilities of frontier models are rapidly accelerating. Our ability to understand, align, and secure them must stay ahead.

__*We will publish a technical report of our learnings in the coming weeks.__