

OpenAI ties large reasoning-distillation campaign to people linked to Moonshot AI

Source copy kept by wheresthe.ai for the story "OpenAI ties large reasoning-distillation campaign to people linked to Moonshot AI" (<https://wheresthe.ai/d/openai-moonshot-distillation-campaign-2026-09/>).

Source: <https://openai.com/index/disrupting-a-coordinated-model-distillation-campaign/>

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

SHA-256 of the saved text: c08066a85c3414d151f41d92335b595a42c2e9f1914684752987513c5cb9cb9a

We recently identified and disrupted a coordinated campaign designed to extract protected reasoning from our models, with the earliest observed activity occurring in the first week of July. This activity is consistent with adversarial distillation: the systematic and unauthorized use of one model's outputs or reasoning to help train, reproduce, or improve another model. Protected reasoning is the model's internal record for working through a task; extracting it can reveal information withheld from the final answer and help others reproduce the model's capabilities.

The operators did not break our encryption, compromise a database, or gain direct access to stored user conversations. Instead, they manipulated model interactions so that protected reasoning could be reproduced in forms visible to the requester in a coordinated, scaled manner that violated our terms of service. This manipulation is not a vulnerability unique to OpenAI's models, and we have shared information about it with industry partners through the Frontier Model Forum in order to strengthen collective defenses against adversarial distillation.

Before publishing, we investigated the scope and potential impact, deployed our own mitigations, and shared with and took feedback from researchers and industry partners to ensure protections against this type of attack are in place. Additional mitigation and investigation work is continuing. We believe sharing what we have learned now will help the broader ecosystem strengthen its defenses.

What we observed

We saw operators attempt to extract protected reasoning in novel ways, including by copying encrypted reasoning from one conversation and asking a model in another conversation to decrypt and transcribe the hidden reasoning content.

Independent security researchers([opens in a new window](#)) also brought related cross-model and conversation-compaction vulnerabilities to our attention through responsible disclosure. We investigated their findings and confirmed that the attack paths they identified were real. Their work helped us understand the broader attack class and accelerate mitigations.

The activity began on July 1, initially at a low volume until we observed high-volume spikes on July 24 and 25 consisting of 16,000 requests¹ using a relevant extraction pattern from over 4,000 users. Further investigation identified related prompt-pattern activity across a cluster of more than 15,000 users, which we fully disrupted by July 28.

The activity evolved over time, reinforcing that adversarial distillation is a broader security challenge that requires layered, adaptive defenses.

Our assessment of attribution

It is unclear whether all operators we observed during the relevant time period originated from a single actor. However, we attribute a core cluster of the activity to individuals associated with Moonshot AI, the developer of Kimi.

Why this matters

Adversarial distillation poses safety and national security risks. Extracted reasoning could be used to train another model without preserving the safeguards applied to the original model's user-facing outputs. At scale, distillation can also accelerate the transfer of advanced capabilities without requiring the same investment in safety. These concerns become heightened as models gain capabilities in dual use domains.

This risk is not unique to OpenAI. As cited above, similar techniques may affect other advanced AI systems,

making this a shared security challenge that requires coordination across the industry.

How we responded

We mitigated this recent distillation campaign through a combination of account enforcement, technical controls, and partner coordination. We banned or restricted fraudulent accounts, strengthened signup and infrastructure controls, and expanded monitoring for related networks.

We also strengthened protections for hidden reasoning across users, workspaces, organizations, and model families. We closed a pathway that allowed someone who already possessed another user's encrypted reasoning to replay it and recover its contents, and added checks to detect and hold streamed output that might expose reasoning. When related activity moved through third-party services, we worked with those providers to identify and disrupt the accounts involved.

Finally, we shared relevant findings through the Frontier Model Forum and appropriate government information-sharing channels so that other frontier developers and public-sector partners could look for similar activity and strengthen their own defenses. Systems that support portable or replayable reasoning artifacts may face related risks.

What comes next

We expect adversarial distillation attempts to become more sophisticated as frontier models improve and as actors look for cheaper ways to mimic their capabilities. Defending against this activity requires layered controls and continual adaptation.

This work is not finished. Partner-hosted deployments need the same protections as first-party services, and tool-output attacks require protections that examine more than ordinary visible text. We are continuing to improve tool defenses, classifier coverage, model refusals, and propagate relevant controls across cloud partners.

Our response will continue to focus on three areas: stronger technical protections against extraction, better detection and enforcement against coordinated campaigns, and deeper threat-information sharing across industry and government.