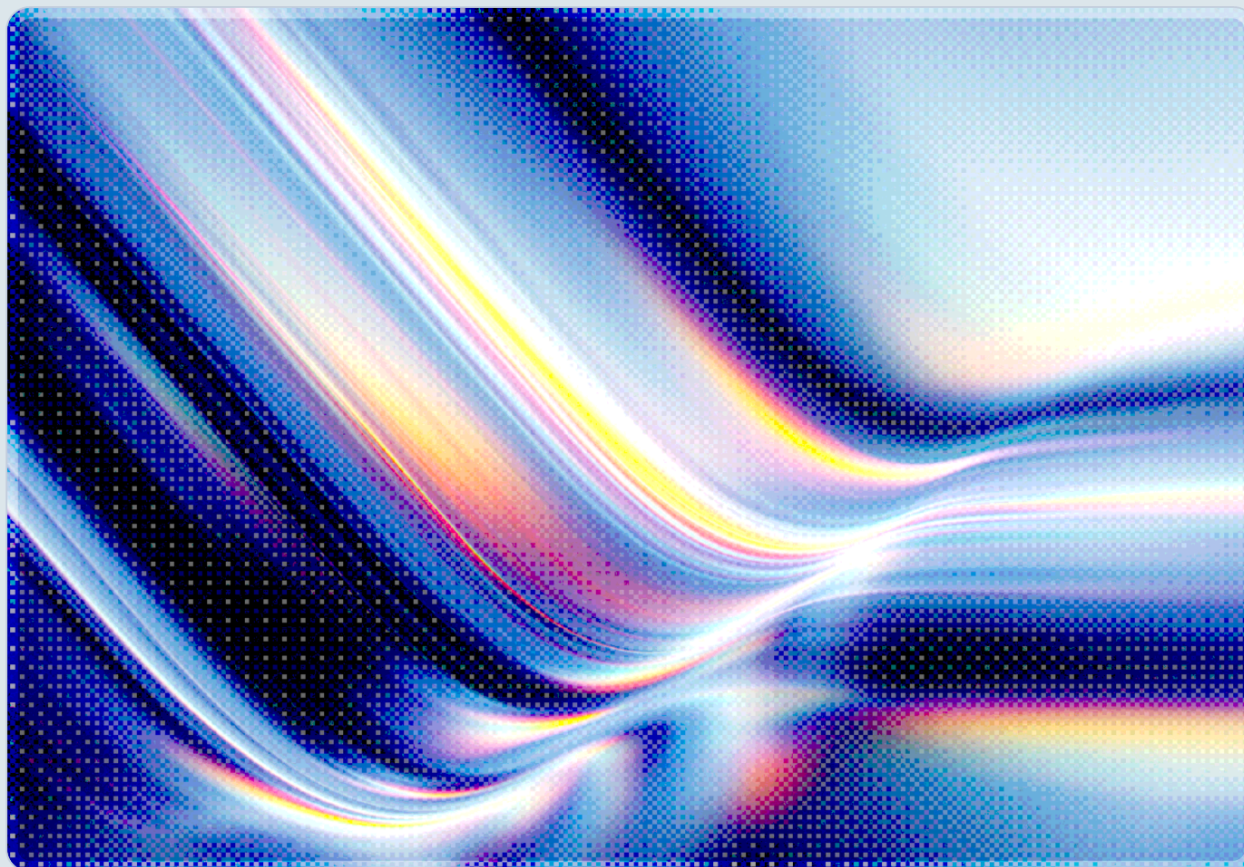


# Reasoning Extraction: Distillation as an Attack on Model Providers

Security Implications and Guidance

2026-10-03

 AI-assisted Rapid Research



CSAI

cloud  
security  
alliance®

**© 2026 Cloud Security Alliance. Some rights reserved.**

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

*This document was generated with AI assistance and has not undergone official CSA review and approval processes.*

---

## Key Takeaways

On September 30, 2026, OpenAI disclosed that it had disrupted a coordinated campaign to extract the protected reasoning of its models, and attributed a "core cluster" of the activity to individuals associated with Moonshot AI [1][2]. According to OpenAI, as relayed by secondary reporting, the technique was not a break of the encryption protecting reasoning traces. Operators reportedly copied encrypted reasoning from one conversation and asked a model in a separate conversation to decrypt and transcribe it, which suggests a session-portability weakness in the reasoning-protection design [1][2][3].

The disclosure fits a broader pattern. Anthropic has publicly accused several China-based laboratories of distillation campaigns against Claude since February 2026 [5][6], and a joint NSA, CISA and FBI advisory in September 2026 reportedly described the activity as systematic [9]. Reasoning traces appear to have become a specific target because they are dense training signal, and providers that hide them are defending a boundary that attackers have been observed probing through chain-of-thought elicitation prompts [5].

Three points matter most for defenders. First, hiding reasoning is a control with its own failure modes, and it needs adversarial testing rather than reliance on encryption alone. Second, detection depends on behavioral and account-network analysis, since individual requests in these campaigns can look like ordinary usage. Third, attribution claims in this area are currently provider assertions that have not been independently verified, and consumers of these reports should treat them accordingly.

## Background

Distillation is a standard machine learning technique in which a smaller or less capable model is trained on the outputs of a stronger one. Providers use it legitimately on their own models, and it is a common route to cheaper deployments. The security concern arises when one party applies the technique to another party's model without authorization, in violation of the provider's terms of service. OpenAI describes this adversarial form as the systematic and unauthorized use of one model's outputs or reasoning to help train, reproduce, or improve another model [1][4].

Anthropic's February 23, 2026 report was an early, detailed public account of this activity at scale. It named DeepSeek, Moonshot AI and MiniMax, and reported more than 16 million exchanges across roughly 24,000 fraudulent accounts, with the traffic attributed by volume to over 150,000 exchanges for DeepSeek, over 3.4 million for Moonshot and over 13 million for MiniMax [5]. Anthropic described commercial proxy services using "hydra cluster" architectures, in which a single network managed more than 20,000 fraudulent accounts at once, and described prompts designed to elicit chain-of-thought output for use as reasoning training data [5]. Later reporting on an Anthropic letter attributed a further campaign to operators tied to Alibaba's Qwen laboratory, described as roughly 28.8 million conversations through about 25,000 fraudulent accounts between April 22 and June 5, 2026 [6][7]. These figures come from press coverage and we did not consult Anthropic's own account of this campaign, so they should be treated as reported rather than confirmed.

A separate allegation concerns Moonshot. Reporting on Anthropic's threat intelligence report in September 2026 states that Moonshot routed almost 300,000 customer requests to Claude over 10 days using 5,380 accounts Anthropic considered fraudulent, presented the answers as Kimi output, and retained some exchanges to extract reasoning for training [8]. That conduct differs from bulk scraping because the extraction is embedded in a live commercial service. We have not located Anthropic's primary text for this claim, and we rely here on secondary reporting.

On September 8, 2026, NSA, CISA and FBI issued advisory AA26-251A. As reported by eSecurity Planet, it names DeepSeek, Moonshot AI, Alibaba, MiniMax, StepFun and Z.AI, describes activity running since at least late 2024, and lists fraudulent account creation, jailbreak prompts that extract chain-of-thought reasoning, and "transfer station" API proxies that evade geographic restrictions as the main methods [9]. Another secondary source gives a different start date of mid-2025 [4], and we did not review the advisory text itself. CSA has already published a research note on that advisory [10]. The OpenAI disclosure adds a technically distinct mechanism to this picture, and it is the focus of the rest of this note.

## Security Analysis

All details of OpenAI's disclosure in this section are second-hand. The OpenAI post [1] could not be retrieved during research, and the dates, counts and quotations below are as reported by The Hacker News [2], CyberScoop [3] and Unite.AI [4], which agree with one another. Where this note characterizes the mechanism, it is our inference from those summaries.

## The OpenAI Campaign

OpenAI reports that low-volume activity began on July 1, 2026. Volume spiked on July 24 and 25, with 16,000 requests matching an extraction pattern from more than 4,000 users. Further investigation found related prompt patterns across a cluster of more than 15,000 users, and OpenAI states the activity was fully disrupted by July 28 [1][2][3][4]. The company says the operators "did not break our encryption, compromise a database, or gain direct access to stored user conversations" [3]. OpenAI also said it was unclear whether all of the observed activity was related, and the core cluster attribution to Moonshot-associated individuals was not accompanied by published technical evidence [2][3]. CyberScoop's reporting does not record a response from Moonshot [3].

Why the technique worked is the instructive part. Where a provider returns encrypted reasoning to the client so that the client can pass it back in later turns, the encrypted blob exists to preserve continuity in a single conversation. The Hacker News summary of OpenAI's report describes the activity as copying encrypted reasoning into a separate conversation, which suggests the traces were usable across sessions [2][3]. If that is the case, an attacker can take a blob produced in one session, present it in another, and prompt the model to transcribe what it contains. The model, which holds the key, performs the decryption on the attacker's behalf. The exact binding between traces and sessions remains to be confirmed against the original post [1].

The lesson may generalize beyond one vendor. Any design in which a model can be induced to emit the contents of a protected artifact has a confused-deputy shape: the cryptography can be sound while the failure lies in what the authorized decryptor can be persuaded to output. If the mechanism is as reported, output filtering, session binding and replay controls should be treated as part of the protection and not as optional hardening around it. CyberScoop reports that outside researchers had described similar weaknesses, which suggests the issue was not unknown to the research community before OpenAI's disclosure [3].

## Why Reasoning Is a Target

Final answers carry limited information about how a model reached them. Reasoning traces carry intermediate steps, error correction and problem decomposition, which in our assessment make them well suited to training a student model. Anthropic reports that the campaigns it observed concentrated on agentic reasoning, tool use and coding, and relied on prompts that elicit chain-of-thought output to generate reasoning data at scale [5]. Providers that hide reasoning may have raised the cost of extraction, which is consistent with attackers investing in bypasses such as the one OpenAI describes, although no cited source reports attacker motivation directly (our inference).

OpenAI also frames the harm as a safety issue. It argues that extracted reasoning could train another model without preserving the safeguards applied to the original model's user-facing outputs, and that capability transfer without matching safety investment becomes more consequential as models gain dual-use capabilities [1][4]. This is an argument from OpenAI, and the degree to which safeguards fail to transfer through distilled data has not, to our knowledge, been demonstrated publicly for these specific campaigns. It is nonetheless a reasonable concern for risk assessment, and it connects distillation to the model-provider threat landscape rather than only to commercial competition.

### Common Attack Patterns Across Providers

The public reports from OpenAI and Anthropic share enough characteristics to support a working taxonomy of how these campaigns operate. The table below summarizes observed techniques and the corresponding detection opportunities. Detection entries marked "(inference)" are CSA's own analysis and were not stated by the cited sources.

| Technique                                                                      | Reported by                                                                   | Detection opportunity                                                                                                                                                   |
|--------------------------------------------------------------------------------|-------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fraudulent account networks, including hydra clusters of thousands of accounts | Anthropic [5]; the advisory reports fraudulent account creation generally [9] | Shared payment methods, synchronized traffic, signup-cluster analysis (inference)                                                                                       |
| Commercial proxies and "transfer stations" that evade regional restrictions    | Anthropic [5], NSA/CISA/FBI [9]                                               | Infrastructure fingerprinting, cloud-provider cooperation (inference)                                                                                                   |
| Chain-of-thought elicitation prompts                                           | Anthropic [5], NSA/CISA/FBI [9]                                               | Classifiers for repetitive, high-volume reasoning-elicitation structure (inference)                                                                                     |
| Replay of encrypted reasoning into a second session                            | OpenAI [1][3]                                                                 | Session binding of traces; detection of outputs that expose reasoning, which OpenAI reports adding [3][4] (inference: monitor for decrypt-and-transcribe prompt shapes) |

| Technique                                                                   | Reported by                | Detection opportunity                                                                    |
|-----------------------------------------------------------------------------|----------------------------|------------------------------------------------------------------------------------------|
| Distillation traffic mixed with unrelated requests                          | Anthropic [5]              | Per-account topical concentration analysis rather than per-request filtering (inference) |
| Alleged relay of a downstream product's customer traffic through a provider | Anthropic, as reported [8] | Volume and geographic anomalies across account groups (inference)                        |

Anthropic describes the detection signature as massive volume concentrated in a few capability areas with highly repetitive prompt structure [5]. That observation has an operational consequence: defenders gain more from analyzing accounts and account networks than from inspecting individual prompts, because each request in a campaign can be unremarkable on its own.

## Attribution and Evidence Quality

Readers should separate three levels of evidence. The first is the technical description of mechanisms, which providers have published in some detail and which is internally consistent across sources. The second is volumetric claims, such as account counts and exchange totals, which only the providers can verify and which we could confirm only through secondary reporting [6][7]. The third is attribution to named organizations, which the providers assert and the named organizations have in several cases not publicly addressed in the sources we reviewed [3]. The named organizations' own positions are not known to us, and a terms-of-service violation is not necessarily a legal one. OpenAI's own statement that it was unclear whether all activity was related, combined with the absence of published attribution evidence, is a reason for hedged language in downstream risk documents [3]. The NSA, CISA and FBI advisory adds government weight to the broader pattern, though the sources we reviewed summarize it secondhand [9].

# Recommendations

## Immediate Actions

Model providers that expose reasoning artifacts to clients should test whether those artifacts can be replayed across sessions, accounts or tenants, and whether the model can be prompted to transcribe their contents. Where replay across sessions is not required for product function, bind artifacts to the originating session or account and reject mismatches. OpenAI reportedly closed pathways allowing replay of encrypted reasoning and added detection for outputs that expose reasoning [3][4]; other providers can use these as a checklist for their own designs.

Providers should also review signup and billing controls for the account-network patterns reported by Anthropic, including shared payment methods and clusters of educational or startup accounts [5]. Enterprises that build products on a frontier API should confirm that their own accounts, keys and integrations cannot be used as a relay by customers or resellers, since the alleged Moonshot arrangement, if accurately reported, shows that relays can serve as an extraction channel [8].

## Short-Term Mitigations

Providers should add account-level behavioral analytics that measure topical concentration, prompt-structure repetition and reasoning-elicitation rates, and should correlate across accounts that share infrastructure or payment instruments. Intelligence sharing among providers, cloud platforms and authorities, which Anthropic recommends, increases the value of these signals because a single provider sees only a fraction of an actor's account fleet [5]. Rate limits and quotas should be tiered by verification level so that newly created or weakly verified accounts cannot produce campaign-scale volume.

Enterprises that consume model APIs should review contractual terms on output use and training, and should record which third-party products embed external models. Where an organization trains or fine-tunes models on outputs from another provider's model, it should confirm that its terms permit this, since the same technique that is an attack in one context is a licensing question in another.

## Strategic Considerations

Providers should treat reasoning protection as a layered control that includes cryptography, session semantics, output monitoring and abuse detection, and should subject it to red-team exercises that focus on getting the model to reveal what it holds. Policy and standards bodies could usefully define a

common vocabulary for adversarial distillation, so that incident reports from different providers can be compared. Finally, risk assessments should avoid reliance on single-source attribution and should record the evidence level of each claim, as discussed above.

## CSA Resource Alignment

CSA's research note on the NSA, CISA and FBI advisory on China's industrial-scale AI distillation campaign [10] is the closest prior work on the campaign pattern. It covers the government assessment of the broader campaign, and this note complements it by examining a specific technical mechanism, the replay of encrypted reasoning, and the evidence-quality questions that follow from provider-issued reports. CSA's *Foundation Model IP Theft: Threat Model for AI Labs* [11] addresses adversary targeting of foundation model intellectual property, which is the wider threat category into which extraction and distillation campaigns fall, and model providers can use it to place the reasoning-extraction scenario within a broader threat model.

The *AICM v1.1 Implementation Guidelines for Model Providers* [12] translate the AI Controls Matrix into obligations for the organizations most exposed to this threat. The recommendations above on session binding, abuse detection and account verification map most directly to the model-provider guidance on securing model access and monitoring for misuse. Providers should assess the reasoning-protection design against these control domains and record the results.

The AI Controls Matrix v1.1 [13] provides the underlying control catalog. Organizations that consume frontier APIs can use it to structure due diligence on provider safeguards and on their own obligations regarding output use.

# References

- [1] OpenAI. "[Disrupting a coordinated model-distillation campaign](#)." OpenAI, September 30, 2026. (The primary page could not be retrieved directly during research; details are as reported by secondary sources [2][3][4].)
- [2] The Hacker News. "[OpenAI Disrupts Reasoning Extraction Campaign Linked to Moonshot AI Associates](#)." The Hacker News, October 2026.
- [3] CyberScoop. "[OpenAI reveals 'novel' encryption bypass used in distillation attack](#)." CyberScoop, September 30, 2026.
- [4] Unite.AI. "[OpenAI Disrupts Coordinated Model-Reasoning Extraction Campaign](#)." Unite.AI, 2026.
- [5] Anthropic. "[Detecting and preventing distillation attacks](#)." Anthropic, February 23, 2026.
- [6] CNBC. "[Anthropic accuses Alibaba of campaign to 'brazenly' and 'illicitly' extract AI capabilities](#)." CNBC, June 24, 2026.
- [7] Let's Data Science. "[Anthropic Accuses Alibaba of 28.8M-Exchange Claude Distillation](#)." Let's Data Science, June 26, 2026.
- [8] RuntimeWire. "[Anthropic says Moonshot secretly served Claude answers through Kimi](#)." RuntimeWire, September 11, 2026.
- [9] eSecurity Planet. "[NSA, FBI, CISA Warn of Industrial-Scale AI Model Distillation Attacks](#)." eSecurity Planet, September 2026 (reporting on advisory AA26-251A, September 8, 2026).
- [10] Cloud Security Alliance. "[CISA Advisory: China's Industrial-Scale AI Distillation Campaign](#)." CSA AI Safety Initiative, September 18, 2026.
- [11] Cloud Security Alliance. "[Foundation Model IP Theft: Threat Model for AI Labs](#)." CSA, 2026.
- [12] Cloud Security Alliance. "[AICMv1.1 Implementation Guidelines for Model Providers \(MP\)](#)." CSA, June 22, 2026.
- [13] Cloud Security Alliance. "[AI Controls Matrix v1.1](#)." CSA, June 22, 2026.