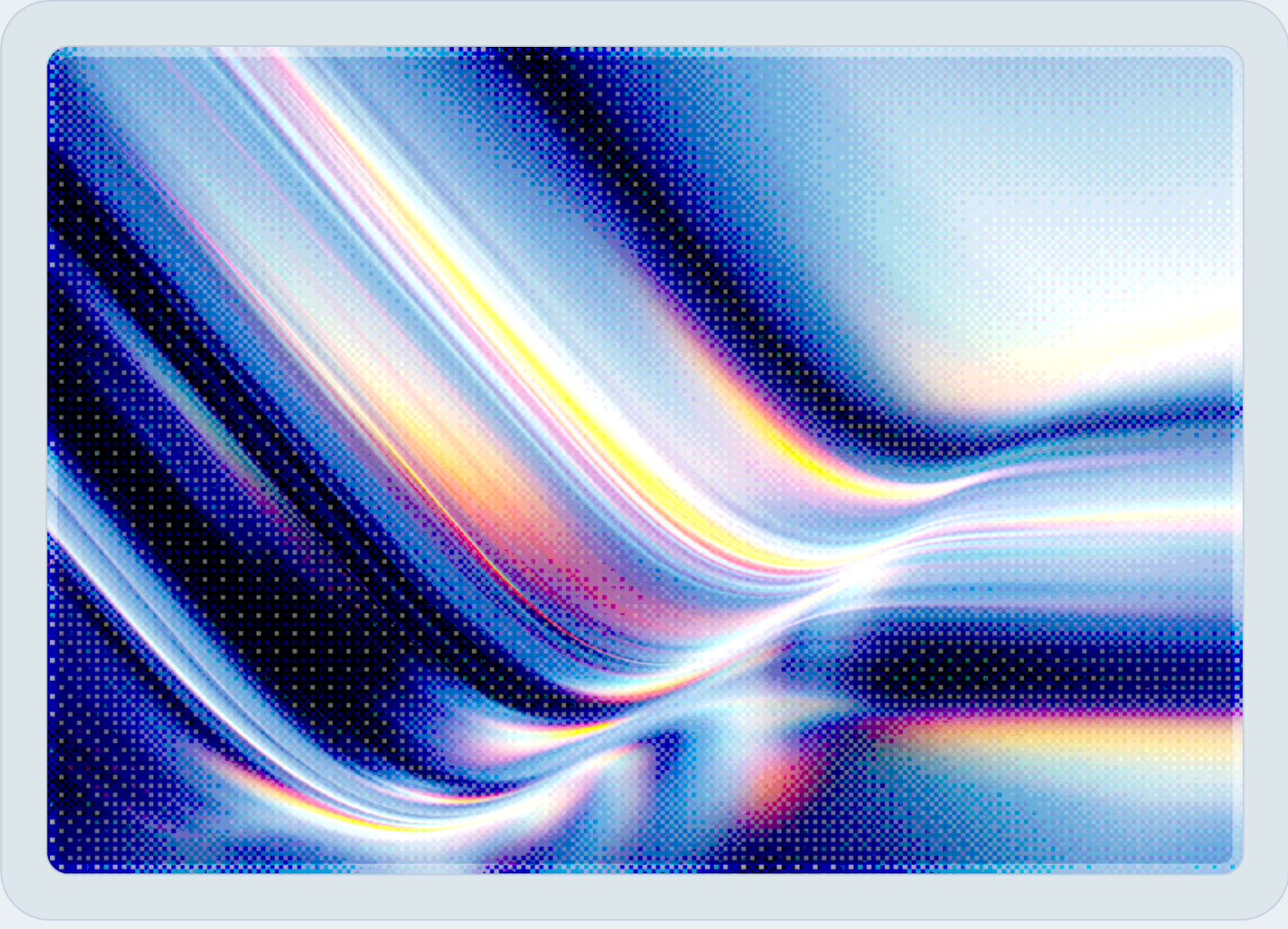
CSAI Foundation | Cloud Security Alliance

NSTM-4: US Policy Response to AI Model Distillation Attacks

Enterprise Implications of the Adversarial Distillation Memorandum

2026-05-02

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- On April 23, 2026, the White House Office of Science and Technology Policy (OSTP) issued Memorandum NSTM-4, "Adversarial Distillation of American AI Models," the first US government policy instrument to formally classify systematic capability extraction from frontier AI systems as a national security threat [1][2].
- Anthropic's February 2026 disclosure documented that three Chinese AI laboratories — DeepSeek, Moonshot AI, and MiniMax — conducted simultaneous but apparently independent extraction campaigns against Claude models using approximately 24,000 fraudulent accounts and, according to contemporaneous media reporting, more than 35,000 API keys, generating over 16 million exchanges over the course of those campaigns, with MiniMax running the largest individual operation [3][4].
- A compounding risk beyond intellectual property theft is safety alignment stripping: models trained purely on the outputs of frontier systems may replicate reasoning capabilities while discarding the safety properties that are harder to transfer through distillation [5][6].
- H.R. 8283, the Deterring American AI Model Theft Act of 2026 (introduced April 2026 and in committee as of this writing), would direct Commerce Department sanctions under IEEPA — including asset freezes and Entity List designations — against entities using improper query-and-copy techniques, with a 180-day assessment and 210-day enforcement timeline if enacted [7][8].
- Enterprises procuring or deploying AI systems must now treat model provenance, vendor attestation practices, and API governance as first-class elements of AI risk programs, as federal compliance provisions and potential supply chain exposure converge around the distillation threat [9][10].

Background

Knowledge distillation is a well-established and legitimate machine learning technique in which a smaller "student" model is trained to approximate the behavior of a larger, more capable "teacher" model. Used responsibly, it enables researchers and engineers to compress expensive frontier models into efficient, deployable versions — a practice widely used by AI laboratories to create on-device and inference-optimized variants of their own systems. The technique is publicly documented, academically studied,

and is a standard technique in practical AI deployment. Nothing in NSTM-4 or the associated legislation targets legitimate distillation; the policy concern is narrow and specific: covert, large-scale extraction campaigns conducted without authorization, in violation of terms of service, and using infrastructure designed to defeat detection [2][11].

The geopolitical context for the memorandum has been developing for more than a year. OpenAI first alleged, in a February 12, 2026 memo to the House Select Committee on China, that accounts associated with DeepSeek employees had deliberately circumvented OpenAI's access controls through obfuscated third-party routers and programmatically extracted model outputs for distillation purposes [12]. Within two weeks, Anthropic published its own disclosure documenting a more extensively instrumented campaign — three distinct Chinese laboratories operating simultaneously, with different query volumes, prompt styles, and evasion techniques [3]. Those February disclosures provided the evidentiary basis that shaped NSTM-4.

DeepSeek released V4 on approximately April 24, 2026, claiming that its new model trails state-of-the-art frontier systems by only three to six months — a gap that would have seemed unlikely given DeepSeek's acknowledged compute constraints two years prior [13][18]. DeepSeek's V4 technical paper describes an approach called On-Policy Distillation (OPD), which draws on outputs from ten separate teacher models [18]. The timing of the White House memorandum closely follows these developments; the juxtaposition of a Chinese lab claiming near-frontier performance through multi-teacher distillation with US labs simultaneously alleging unauthorized extraction of their model outputs appears to have created the political and technical conditions for a formal government response.

Security Analysis

The Adversarial Distillation Technique

Adversarial distillation is a systematic process in which an attacker submits a large volume of carefully engineered queries to a target frontier model, collects the full responses, and uses those input-output pairs as training data for a competing model. The attack exploits an asymmetry inherent to AI service delivery: the outputs of a frontier model contain encoded representations of its capabilities, reasoning patterns, and knowledge, and those outputs are the only artifact the API must return to remain functional. Unlike traditional IP theft, which requires compromising source code or model weights, adversarial distillation extracts value through the public API interface, making conventional perimeter defenses insufficient [5][14].

At scale, attackers use several techniques to maximize extraction efficiency and evade detection. Prompt diversity campaigns submit systematically varied queries across domains, reasoning formats, and instruction styles to probe the full breadth of the target model's capability surface. Chain-of-thought elicitation specifically targets responses that expose the model's reasoning process, which is particularly valuable for training student models in advanced inference behaviors. Evasion infrastructure — including commercial proxy services, rotating residential IP addresses, and coordinated multi-account networks — spreads traffic across thousands of apparent users to defeat per-key and per-IP rate limits [3][6]. In the Anthropic campaigns, at least one proxy network simultaneously managed more than 20,000 fraudulent accounts, interleaving distillation traffic with unrelated API calls to reduce the signal-to-noise ratio in usage telemetry [3].

Observed Industrial-Scale Campaigns

The documented scale of the February 2026 campaigns is notable on its own terms. Anthropic's disclosure identified three concurrent, separately operated campaigns generating a collective 16 million exchanges [3]. Contemporaneous reporting attributed this activity to more than 35,000 API keys used over approximately seven days, though those figures originate in sources not fully accessible for independent verification [4]. MiniMax ran the largest individual operation, accounting for over 13 million of those exchanges. DeepSeek and Moonshot AI operated distinct campaigns with different behavioral signatures, suggesting separate engineering teams and varying extraction objectives rather than a single coordinated operation.

OpenAI's concurrent disclosure adds a distinct dimension: the company reported observing code written by individuals associated with DeepSeek explicitly designed to route extraction queries through third-party relay infrastructure to obscure the origin and defeat account-level detection [12]. This suggests that by early 2026, adversarial distillation had matured from opportunistic API abuse into a structured discipline, with purpose-built tooling and operational security practices suggesting a level of operational discipline previously associated with state-affiliated intrusion campaigns. Google's Threat Intelligence Group separately noted distillation experimentation as a growing pattern across the adversarial AI threat landscape, with frontier capabilities from multiple US providers targeted simultaneously [15].

Safety Alignment Stripping

Perhaps the least-discussed consequence of adversarial distillation is what is lost in the transfer, rather than what is gained. Safety properties — the behavioral constraints that cause frontier models to decline harmful requests, flag dangerous instructions, and maintain consistent ethical guardrails — are emergent characteristics arising from reinforcement learning from human feedback, constitutional AI training, and

other alignment techniques applied after pre-training. These properties are encoded in the model's learned preferences and refusal behaviors rather than in discrete knowledge representations, and as such are not straightforwardly captured in training data derived from API outputs alone [5][6].

Current evidence and expert commentary suggest these properties are harder to reliably transfer through distillation than factual knowledge or reasoning capability, because the alignment signal that shapes a frontier model's behavior is not fully represented in the surface-level outputs an extraction campaign collects [5][6]. A student model trained exclusively on the query-output pairs of an extraction campaign will tend to acquire the capability surface of the frontier model while reproducing its safety behavior inconsistently or not at all. The extracted training data reflects what the frontier model said, but not the full alignment signal that shaped why it said it in that way.

The practical result is that illicitly distilled models may respond to sensitive requests in ways that frontier models would refuse, behave inconsistently when tested against known jailbreak patterns, and lack the documented alignment properties and behavioral attestations that enterprise procurement and regulatory compliance increasingly require. This is a supply chain risk, not merely an IP dispute: if organizations deploy commercial AI products whose underlying model was trained on extraction data from a frontier system, they may be acquiring capability without the safety properties they assumed were part of the package [9][10].

National Security and Enterprise Exposure

NSTM-4 frames the threat primarily in terms of national security and competitive advantage — adversaries circumventing billions of dollars in research investment by extracting capabilities through the API layer [1][2][17]. But the memorandum also has direct operational implications for enterprises that are neither AI laboratories nor federal agencies. The near-term directives in NSTM-4 require agencies to update vendor contracts with new attestation requirements, mandating that AI providers document their controls against adversarial extraction and attest to the integrity of their training pipelines [1][2]. Federal contractors and organizations participating in federal AI programs will face these requirements first, but the compliance pattern is likely to diffuse across procurement standards more broadly, particularly as analysts have projected that high-risk supply chain compliance provisions anticipated for August 2026 may introduce additional attestation requirements for AI system components.

Enterprises that rely on commercial AI APIs face a parallel exposure. If a vendor's model was partially trained using distillation data extracted from a competitor's system, the enterprise inherits whatever safety alignment gaps or IP encumbrance that training data carried. Organizations currently lack standardized tools for assessing model provenance at procurement time; vendor attestation, contractual representations, and third-party audit are the primary available mechanisms. The InformationWeek

analysis of distillation risks to enterprise AI supply chains notes that CIOs must treat model provenance as analogous to software bill-of-materials (SBOM) — a supply chain integrity concern requiring verifiable documentation of training data lineage, not just functional capability testing [9].

The threat is not limited to procurement-side exposure. Enterprises operating their own AI APIs — either as AI providers to customers or as organizations exposing internal models via API for business applications — may themselves be targets of extraction campaigns. Organizations with valuable proprietary AI capabilities developed through fine-tuning, retrieval-augmented generation customization, or domain-specific training face a structurally similar extraction risk, though the economic and strategic incentive for adversaries differs substantially from that faced by frontier model providers [5].

Recommendations

Immediate Actions

Enterprises should immediately audit API usage patterns on all AI services they operate or expose. Establish baseline traffic profiles — query volume per key, query diversity, request timing distributions, prompt length distributions — and configure alerting for deviations that are consistent with systematic extraction: high volumes from a single key or IP, unusual uniformity in prompt structure, or diversity of topics far exceeding normal use-case variance. Rate limiting is a necessary but insufficient first response; per-key quotas can be defeated by distributed multi-account campaigns, so behavioral analytics must complement volumetric controls.

Organizations that have deployed AI APIs without identity verification controls should evaluate the feasibility of requiring stronger account authentication — phone verification, payment instrument verification, or organizational attestation — for API access. Anthropic's post-incident response included expanded account graphing and stricter identity verification as primary structural defenses [3][4]. For enterprises operating third-party AI APIs in internal systems, review the API key management posture: keys shared across applications or stored without rotation policies increase the attack surface and complicate forensic attribution if extraction activity is detected.

Short-Term Mitigations

As a near-term priority, organizations procuring commercial AI services should request vendor attestation regarding model training data provenance and distillation controls. This is particularly important for organizations subject to federal contracting requirements or operating under compliance

frameworks that require vendor risk documentation. Specifically, ask vendors whether they have implemented: behavioral anomaly detection on API traffic, watermarking or output perturbation controls to enable distillation attribution, account graphing for coordinated abuse detection, and a documented incident response process for distillation campaigns. Vendors that cannot provide substantive answers to these questions represent elevated supply chain risk.

Legal and compliance teams should track H.R. 8283's progress through the House Committee on Foreign Affairs and prepare to assess vendor relationships with entities that may fall within its scope. The bill's 180-day assessment timeline means that a potential Entity List designation tranche could arrive in late 2026 or early 2027 if the legislation is enacted. Vendor contracts that currently lack provisions addressing AI IP integrity and adversarial training data should be flagged for amendment at the next renewal cycle. Organizations participating in federal AI programs should anticipate that NSTM-4's attestation requirements will be incorporated into agency procurement standards, and proactively request that vendors confirm compliance with the memorandum's near-term directives.

Strategic Considerations

The longer-term governance question raised by NSTM-4 and its companion legislation is whether AI model provenance will become a regulated compliance requirement analogous to software supply chain integrity. The trajectory from Executive Order 14028 on software supply chain security to mandatory SBOM requirements for federal contractors is an instructive precedent. Organizations that treat AI model provenance documentation as a strategic investment now — rather than a reactive compliance exercise later — are likely to be better positioned if and as compliance requirements develop in this area [16]. This means incorporating model training data attestation, safety alignment documentation, and distillation-risk disclosures into AI vendor due-diligence frameworks.

Security teams should also assess internal AI capabilities for distillation-target risk. Any proprietary model made accessible via API — whether a fine-tuned language model, a specialized classification system, or a custom agent — should be evaluated for its value as an extraction target. Models representing significant investment in domain-specific training or alignment work are the most attractive candidates for adversarial distillation. Architectural controls such as output perturbation, query watermarking, and rate structures calibrated to discourage bulk extraction can reduce exposure without degrading legitimate use. The OSTP memorandum's near-term directives include integrating watermarking and perturbation into production models as a mandatory step for federal AI deployments [1][2], providing a useful baseline for enterprise security teams evaluating analogous controls.

CSA Resource Alignment

The adversarial distillation threat maps directly onto several domains of CSA's AI Controls Matrix (AICM). The Model Security domain addresses controls governing the integrity of model training pipelines, training data provenance, and defenses against adversarial manipulation of the model development lifecycle. NSTM-4's attestation requirements and the vendor compliance implications described above align with AICM's supply chain security domain, which requires organizations to document the provenance of AI components and establish contractual controls over AI vendor security practices. The AICM supersedes the earlier Cloud Controls Matrix (CCM) for AI-specific governance and should serve as the primary mapping framework for organizations building compliance programs around the distillation threat.

CSA's MAESTRO threat modeling framework for agentic AI systems provides a structured methodology for modeling distillation as an adversarial threat to AI supply chains. MAESTRO's threat decomposition approach is applicable to both the provider side — modeling how an adversary extracts capabilities from an exposed model — and the consumer side — modeling how an organization's deployed systems might be compromised by a vendor whose model training integrity is in question. Security teams using MAESTRO should include adversarial distillation in their threat libraries alongside prompt injection and model poisoning.

The CSA STAR (Security Trust Assurance and Risk) registry and its associated cloud controls provide the vendor assessment framework most immediately applicable to the attestation requirements emerging from NSTM-4. Organizations should request that AI vendors complete STAR attestations that address distillation risk controls, and treat absence of such documentation as a risk signal in procurement decisions. CSA's Zero Trust guidance applies to API access governance: the principle of least privilege applied to AI API access means issuing scoped, short-lived credentials tied to specific use cases rather than long-lived, broadly scoped API keys that create an attractive target surface for bulk extraction campaigns.

References

[1] Legal Wire. "OSTP Issues Memorandum Alleging Foreign Entities Are Conducting Coordinated Campaigns to Distill US Frontier AI Systems." The Legal Wire, April 2026.

[2] Nextgov/FCW. "White House Accuses China of 'Deliberate, Industrial-Scale Campaigns' to Steal US AI Models." Nextgov/FCW, April 23, 2026.

[3] Anthropic. "Detecting and Preventing Distillation Attacks." Anthropic News, February 23, 2026.

[4] CNBC. "Anthropic Joins OpenAI in Flagging 'Industrial-Scale' Distillation Campaigns by Chinese AI Firms." CNBC, February 24, 2026.

[5] MindStudio. "AI Model Distillation Attacks: What They Are and Why They Matter." MindStudio Blog, 2026.

[6] Metora/TIE. "Countering Industrial-Scale AI Model Distillation and Theft." Technology Intelligence Exchange, 2026.

[7] Congress.gov. "H.R. 8283 — Deterring American AI Model Theft Act of 2026." Congress.gov, 119th Congress.

[8] GovInfo. "H.R. 8283 (IH) — Deterring American AI Model Theft Act of 2026." GovInfo, 2026.

[9] InformationWeek. "Distillation Attacks Expose Hidden Risk in Enterprise AI Supply Chain." InformationWeek, 2026.

[10] Corporate Compliance Insights. "2026 Operational Guide to Cybersecurity, AI Governance & Emerging Risks." Corporate Compliance Insights, 2026.

[11] NATO Lambert (Substack). "On Policy Targeting Distillation 'Attacks'." natolambert overflow, 2026.

[12] Bloomberg. "OpenAI Accuses China's DeepSeek of Distilling US AI Models to Gain an Edge." Bloomberg, February 12, 2026.

[13] Euronews. "China's DeepSeek Releases New AI Model V4: Everything to Know as the AI Race Speeds Up." Euronews, April 24, 2026.

[14] IAPS. "AI Distillation Attacks: The Case for Targeted Government Intervention." Institute for AI Policy and Strategy, 2026. (URL unavailable as of 2026-05-02; paper may have moved or been archived.)

[15] Google Cloud. "GTIG AI Threat Tracker: Distillation, Experimentation, and (Continued) Integration of AI for Adversarial Use." Google Cloud Blog, 2026.

[16] Just Security. "The Case for Imposing Costs on China's AI Distillation Campaigns." Just Security, 2026.

[17] Defense One. "China Has 'Deliberate, Industrial-Scale Campaigns' to Steal US AI Models, White House Says." Defense One, April 23, 2026.

[18] Asia Times. "US Sounds Alarm on China's AI Distillation as DeepSeek V4 Debuts." Asia Times, April 2026.