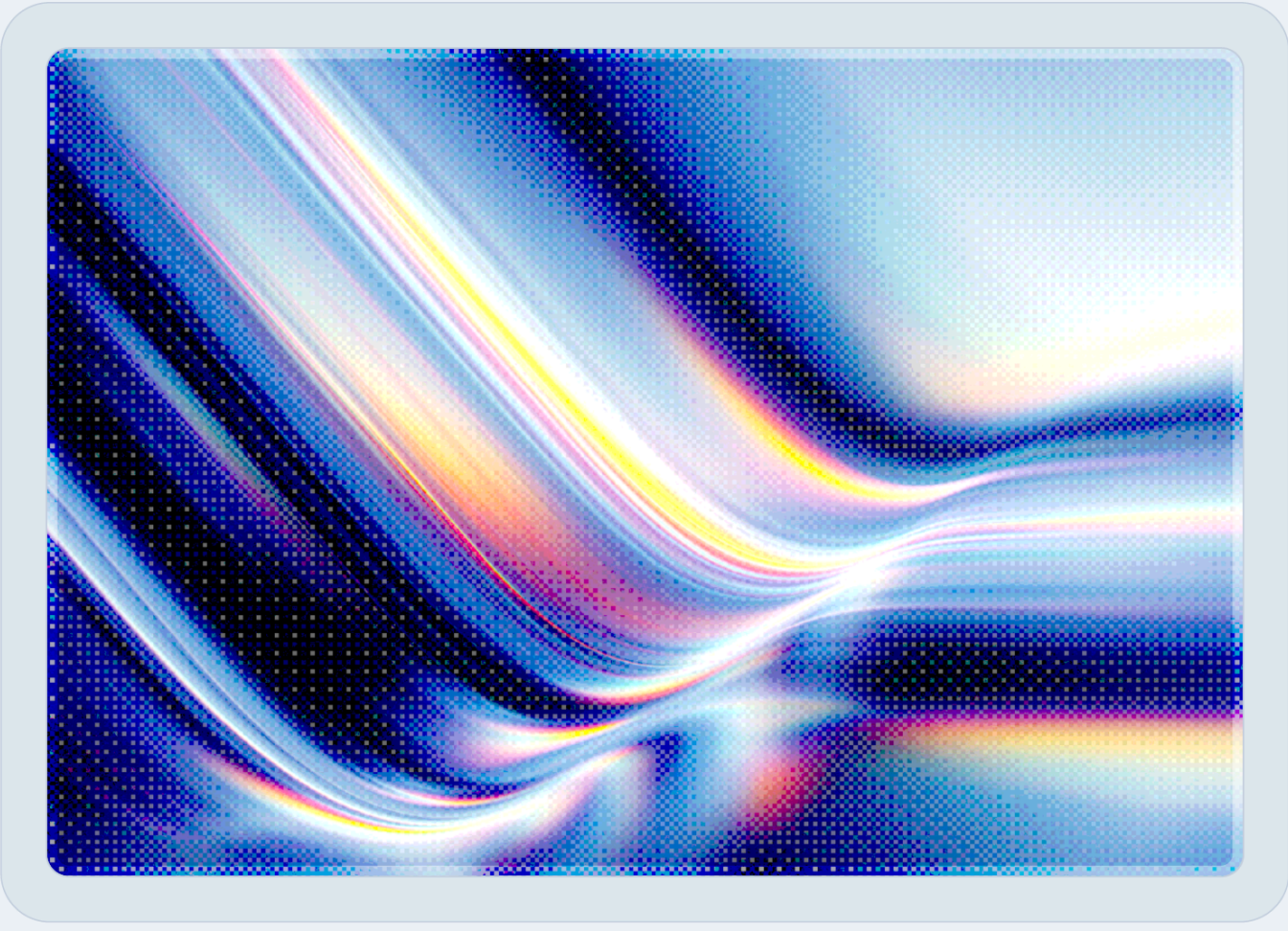
CSAI Foundation | Cloud Security Alliance

Industrial-Scale Model Distillation as a National Security Problem

Assessing the NSA-CISA-FBI Advisory on China-Based AI Extraction Campaigns

2026-09-10

AI-assisted Rapid Research



CSAI | cloud security alliance®

Table of Contents

Executive Summary

On September 8, 2026, the National Security Agency, the Cybersecurity and Infrastructure Security Agency, and the Federal Bureau of Investigation issued a joint cybersecurity advisory, tracked as AA26-251A, asserting that six China-based artificial intelligence companies have conducted "industrial-scale knowledge distillation campaigns" against American frontier AI models since at least late 2024 [1][2]. The advisory names DeepSeek, Moonshot AI, Alibaba, MiniMax, StepFun, and Z.AI, and it describes extraction activity against models from Anthropic, OpenAI, Google, and xAI totaling, in the agencies' words, billions of tokens across millions of exchanges [1][3]. What distinguishes this advisory from ordinary competitive-intelligence reporting is its framing: the three agencies characterize distillation not as a supplementary technique but as "the core—not merely a supplement" of the named companies' AI development strategy, and they assess that the activity is occurring "likely with Chinese government awareness" [1][4]. That framing converts a long-simmering commercial and intellectual-property dispute into a matter the U.S. government is prepared to treat with the same institutional machinery it applies to state-sponsored cyber intrusion.

This paper examines what the advisory actually alleges, how its technical claims compare against the months of corporate threat reporting that preceded it, and what the shift from private complaint to joint federal advisory signals for AI providers, enterprises, and policymakers. The advisory did not emerge from nowhere. Anthropic reported in February 2026 that three of the same six companies had generated more than sixteen million suspicious exchanges with Claude from roughly twenty-four thousand fraudulently created accounts [8]. Google's Threat Intelligence Group has since documented a related pattern against Gemini, including a reasoning-trace extraction campaign — identified through more than one hundred thousand prompts spanning multiple languages — that targeted the model's underlying chain-of-thought output [10]. The government advisory synthesizes this pattern, adds tactic and technique detail mapped to the MITRE ATLAS framework, and issues it as an interagency assessment carrying the institutional weight of national security policy rather than corporate self-interest.

The paper argues that three developments matter more than the headline token and exchange figures the advisory reports. First, the government has adopted the vocabulary of espionage and supply-chain compromise for a technique — distillation — that is also a standard, legitimate machine learning practice used throughout the industry, including by the very companies filing the complaints. That ambiguity is not incidental; it is the crux of the dispute, and China's Ministry of Commerce has seized on it directly, calling distillation "a neutral technical means used by global model companies including U.S. firms" and warning of "resolute countermeasures" if the advisory is used to justify new restrictions [11]. Second, the advisory's recommended mitigations — quietly degrading suspected accounts rather than banning them,

sharing behavioral indicators across competing providers, and applying differential privacy to API outputs — ask frontier AI companies to build defensive infrastructure that, in our assessment, existed at only a nascent stage as an operational discipline before the corporate disclosures of early 2026. Third, this advisory arrives in a policy environment already shaped by the first-ever suspension of a commercial frontier model under export control authority, meaning distillation enforcement will not be evaluated in isolation but as part of a broader pattern of the U.S. government asserting direct authority over frontier model access and use [14].

For CISOs and AI governance leaders, the practical implications are immediate: API terms of service, usage monitoring, and incident response plans built for conventional account abuse were not designed for adversaries capable of routing tens of thousands of coordinated accounts through gray-market proxy infrastructure, and, in our assessment, few organizations yet have the telemetry, cross-provider information-sharing relationships, or legal clarity needed to detect or respond to this pattern.

Introduction and Background

Knowledge distillation is not new, and it is not, on its own, malicious. First formalized as a machine learning technique in the mid-2010s [20], distillation trains a smaller "student" model to reproduce the outputs of a larger "teacher" model, typically by having the student learn from the teacher's probability distributions over possible answers rather than from raw labeled data alone. Model providers use distillation internally to compress large models into smaller, cheaper variants for deployment; academic researchers use it to study what large models have learned; and much of the open-source AI ecosystem depends on techniques that are, structurally, forms of distillation. The technique's legitimacy is precisely what makes the current dispute difficult to adjudicate cleanly, and it is the argument China's government has made in public response to the U.S. advisory [11].

What changed over the past eighteen months is scale, intent, and method. Rather than a researcher fine-tuning an open model on a modest set of teacher outputs for a defined research purpose, the pattern the NSA, CISA, and FBI describe involves organized programs routing millions of queries through fraudulently created accounts, gray-market proxy services, and third-party API aggregators specifically to harvest a frontier model's outputs at industrial volume, while evading the rate limits, terms of service, and geographic restrictions each provider had put in place [1][5]. The advisory places the origin of this pattern at "at least late 2024," roughly coinciding with DeepSeek's emergence as a global story following the January 2025 release of its R1 reasoning model, which the company claimed to have trained for approximately $5.6 million — a figure the advisory explicitly disputes as misleading because it excludes the cost of the distilled training data DeepSeek is alleged to have extracted from Western frontier models [1][3].

The dispute did not begin with the government. Anthropic was the first to make the accusation publicly and specifically, reporting in February 2026 that DeepSeek, Moonshot AI, and MiniMax had collectively generated more than sixteen million exchanges with Claude from approximately twenty-four thousand fraudulently created accounts, with MiniMax alone responsible for more than thirteen million of those exchanges [8]. Anthropic warned at the time that illicitly distilled models could lack the safety guardrails built into the original systems and could enable "authoritarian governments to deploy frontier AI for offensive cyber operations, disinformation campaigns, and mass surveillance" [8]. Google's Threat Intelligence Group has since reported a comparable pattern, observing distillation activity against Gemini that includes systematic extraction of the model's reasoning traces, using proxy infrastructure and rotating credentials to obscure its origin [10]. By September 2026, reporting indicated Anthropic's own investigation had extended into gray-market forums where access to distillation-ready accounts and API pathways was reportedly being bought and sold [9]. The government advisory is, in this sense, less a new discovery than a formal, interagency ratification of a pattern three of the largest U.S. AI labs had already been separately documenting and separately escalating for roughly seven months.

This whitepaper is organized around three questions. First, what does the advisory actually claim, and how well-supported are those claims by the corporate threat intelligence that preceded it? Second, what does the government's choice to frame distillation as a national security threat — rather than leaving it to civil litigation, trade complaints, or bilateral diplomacy — signal about how the U.S. intends to govern the AI supply chain going forward? Third, what should AI providers, enterprises, and CISOs do in response, given that the advisory's own recommended mitigations require capabilities many organizations do not yet have?

The Advisory: Claims, Companies, and Scope

The joint advisory names six China-based companies and describes each as running a distinct extraction campaign against a distinct set of Western models, summarized in Table 1. The agencies were explicit that they consider distillation to be more than a supplementary research technique for these companies: it is, in the advisory's language, "the core—not merely a supplement" of their AI development strategy, extracting not just raw text outputs but specialized capabilities including legal reasoning, chain-of-thought and agentic function execution, software engineering proficiency, and fine-tuned dialogue behavior [1][5].

Table 1: Companies and Alleged Activity per the Joint Advisory

Chinese Company	Approximate Campaign Window	Targeted U.S. Models	Primary Capabilities Extracted
DeepSeek	Late 2024 – mid-2025	Claude 3.7/Sonnet 4/Opus 4.1, GPT-4o, GPT-5	Reasoning, chain-of-thought (R1, V3 development)
Moonshot AI	Mid-2025 onward	Claude (including Fable 5), GPT-4o	Fine-tuning, agentic and software-engineering functions (Kimi-K2/K3)
Alibaba	Late 2025	Claude, GPT-5 variants	Software engineering, coding optimization
MiniMax	Ongoing since 2025	Claude, Gemini, GPT	Dialogue, reasoning, rapid model-release targeting
StepFun	Late 2025 – early 2026	Claude Opus/Sonnet/Haiku, GPT-5 series	Reasoning and specialized task optimization
Z.AI	Mid-2026	GPT-5.5, Claude Opus 4.8	Bulk token extraction, general capability transfer

Source: compiled from the joint advisory and contemporaneous reporting [1][5][6][7].

The scale figures in the advisory are directional rather than precise. The agencies describe "billions of tokens across millions of exchanges/requests" extracted over the campaign period, and they characterize per-domain campaigns as ranging from thousands to millions of queries, but the public version of the advisory does not provide a single reconciled total, a breakdown by company, or independently auditable evidence [1]. This is broadly consistent with the format CISA has used for other joint cybersecurity advisories, which typically convey indicators and tactics that defenders can act on rather than serve as forensic reports suitable for legal proceedings. Readers evaluating the advisory's claims should weigh the specific, numbered figures that individual companies have separately published — such as Anthropic's sixteen-million-exchange, twenty-four-thousand-account count — more heavily than the advisory's aggregate language, which synthesizes multiple providers' internal telemetry that the public cannot independently verify [1][8].

The advisory's most operationally useful content is not the attribution itself but the detailed tactics, techniques, and procedures it documents, several of which the agencies characterize as novel. These are examined in the next section.

Anatomy of an Industrial-Scale Distillation Campaign

The advisory maps the observed activity to MITRE ATLAS, the adversarial threat framework purpose-built for machine learning systems, citing techniques including AI Model Inference API Access (AML.T0040), LLM Prompt Injection (AML.T0051), LLM Jailbreak (AML.T0054), and Exfiltration via AI Inference API (AML.T0024.002) [1][16]. Grounding the advisory in an established framework rather than ad hoc description is a meaningful methodological choice: it allows defenders to reuse existing ATLAS-aligned detection and mitigation catalogs rather than build a bespoke response to a single advisory, and it signals that the U.S. government now treats large-scale model extraction as belonging to the same analytical category as other adversarial machine learning threats rather than as a purely commercial or intellectual-property matter.

Three layers of technique recur across the companies named in the advisory. The first is access obfuscation. The agencies describe "transfer stations" — gray-market proxy networks, some reportedly advertised on Chinese consumer marketplaces such as Taobao and Xianyu, that route requests through obfuscated accounts to bypass the geographic and usage restrictions providers have put on frontier model access [1][5]. Operators reportedly manage tens of thousands of fraudulent accounts simultaneously, distributing queries across native APIs, cloud-hosted endpoints, and third-party API aggregators so that no single access pathway shows the volume or pattern that would otherwise trigger a provider's abuse detection [1][5].

The second layer is extraction technique proper. Beyond simply querying a model for its outputs, the advisory describes systematic chain-of-thought extraction — using jailbreak-style prompts and prompt injection to force a model to expose its internal reasoning steps rather than just its final answer, which is disproportionately valuable to a company trying to replicate a frontier model's reasoning capability rather than its surface-level responses [1][5]. The advisory also describes quality-evaluation frameworks the Chinese companies allegedly use to detect when a target provider has deployed defensive countermeasures, such as degraded or randomized outputs, and to route around them, and it documents at least one instance in which a company redirected its extraction activity to a newly released Claude model within twenty-four hours of that model's public launch [1].

The third layer is operational resilience. The advisory identifies automated failover between access pathways when one is blocked, metadata sanitization performed at the infrastructure layer to strip identifying signals from requests, and continuous optimization of account and subscription usage to extract maximum output per dollar spent on premium API access [1][5]. Table 2 summarizes the technique categories against their ATLAS mapping and the defensive countermeasures the advisory and complementary NIST guidance recommend.

Table 2: Distillation Campaign Techniques, ATLAS Mapping, and Recommended Countermeasures

Technique Category	MITRE ATLAS Reference	Advisory-Recommended Countermeasure
Access obfuscation (proxies, fraudulent accounts, aggregators)	AI Model Inference API Access (AML.T0040)	Behavioral monitoring of subscription-to-usage ratios and cross-account correlation
Chain-of-thought / reasoning extraction	LLM Prompt Injection (AML.T0051); LLM Jailbreak (AML.T0054)	Output obfuscation; reduced reasoning-trace fidelity for suspected distillation attempts
Bulk output harvesting	Exfiltration via AI Inference API (AML.T0024.002)	Query volume and rate limiting; differential privacy with calibrated output noise
Countermeasure evasion and failover	Not separately coded; treated as operational TTP	Cross-organization intelligence sharing; telemetry correlation across providers and cloud platforms

Source: compiled from the joint advisory's MITRE ATLAS mappings and NIST AI 100-2e2025 recommendations [1][16][17].

The advisory's most consequential — and most contested — recommendation is its guidance on response. Rather than instructing providers simply to detect and ban suspected distillation accounts, the agencies recommend deploying "subtly altered responses for suspected malicious distillation attempts," including reduced reasoning depth or alternative reasoning pathways, and explicitly advise against notifying the suspected account holder that its responses have been altered [1]. This is a deliberate deception strategy applied at the product layer, and it asks frontier AI companies to build and maintain

classifiers capable of distinguishing distillation activity from ordinary heavy usage with enough confidence to justify silently degrading service — a capability that carries its own risk of false positives against legitimate high-volume customers.

From Corporate Alarm to Government Assessment

The interagency advisory did not create the distillation narrative; it formalized one that had already been building for roughly seven months through independent corporate threat reporting, and understanding that progression matters for judging how much new information the advisory actually contributes.

Anthropic's February 2026 disclosure was the first to attach specific, named companies and enumerated figures to the pattern, reporting sixteen million-plus exchanges from roughly twenty-four thousand fraudulently created accounts across DeepSeek, Moonshot AI, and MiniMax, with MiniMax responsible for the largest single share [8]. That disclosure drew a sharp distinction Anthropic has since repeated: that distillation is "a widely used and legitimate training method" but that the specific pattern of account fraud, volume, and evasion it observed indicated use "for illicit purposes" — a framing that anticipates, and largely mirrors, the government advisory's own core distinction between distillation-as-technique and distillation-as-systematic-extraction [1][8]. By September 2026, contemporaneous reporting indicated Anthropic's investigation had extended into gray-market forums where distillation-ready account access was reportedly being traded, suggesting the underlying access-obfuscation infrastructure the government advisory describes has itself become a commercial service layer rather than a set of one-off workarounds [9].

Google's Threat Intelligence Group has independently documented a related pattern against Gemini. Its threat-tracking report describes a "Reasoning Trace Coercion" case study in which attackers issued more than one hundred thousand prompts across multiple languages specifically to force exposure of the model's underlying chain-of-thought output — evidence that the reasoning-extraction technique the advisory describes for Claude and GPT models is not confined to a single provider [10]. Google's reporting, like Anthropic's, describes proxy infrastructure and credential rotation as an evasion mechanism, which is consistent with the "transfer station" and metadata-sanitization techniques the joint advisory later formalized [1][10].

Three of the four major U.S. frontier labs — Anthropic, OpenAI, and Google — began sharing threat intelligence on distillation activity through the Frontier Model Forum in April 2026, an industry body founded alongside Microsoft in 2023 to coordinate frontier-model safety practices, though reporting on the arrangement indicates the depth of that cooperation remains constrained by uncertainty over what antitrust rules permit competitors to share with one another [18]. This cross-provider coordination is

itself notable: it suggests that no single provider's telemetry was sufficient to establish the pattern the government advisory now describes, and that the interagency assessment likely draws, directly or indirectly, on intelligence that was first assembled cooperatively among competing commercial labs before it reached a classified or law-enforcement channel. That sequencing — corporate detection, cross-provider sharing, then government ratification — is suggestive, though this single case does not by itself establish a general rule, of how future AI supply-chain threats may surface: the private sector may prove to be the leading indicator for this category of threat, with federal advisories arriving as a lagging, aggregating confirmation rather than a first alarm.

The National Security Reframing

The substantive shift the joint advisory represents is not technical but categorical. Model distillation via API abuse has been technically possible, and commercially alleged, for as long as frontier models have been offered as a paid service. What is new is the U.S. government's decision to treat it using the institutional apparatus normally reserved for nation-state cyber intrusion: a joint advisory co-signed by the NSA, CISA, and FBI, structured around MITRE ATLAS tactic mappings, indicators of compromise, and TTP enumeration in the same format used for ransomware groups and state-sponsored intrusion sets [1].

That framing carries three implications that extend beyond the six companies named. First, it establishes model outputs — not just model weights, training data, or infrastructure — as a protected category of intellectual property whose systematic extraction the U.S. government is willing to characterize as a national security concern rather than purely a private contractual or copyright dispute between a provider and a user who violated its terms of service. This is a meaningful expansion. Export control law has increasingly extended beyond physical goods and technical data to cover model weights and training compute above defined thresholds in recent U.S. policy actions; treating the outputs of authorized API access as a controllable, protectable asset is a substantially newer and more contested legal terrain [12]. Second, the advisory's assessment that this activity is occurring "likely with Chinese government awareness" imports the vocabulary of state attribution into a dispute whose underlying facts remain, on the public record, genuinely contested [1][4]. Whether that awareness rises to direction, tolerance, or simple non-interference is a distinction the advisory does not resolve, and it is precisely the distinction China's Ministry of Commerce disputed in its response, which characterized distillation as ordinary technical practice and rejected the suggestion of state involvement or malicious intent outright [11].

Third, and most consequentially for enterprise planning, this advisory does not arrive in a vacuum. It follows, by less than three months, the U.S. Commerce Department's unprecedented use of export control authority to suspend global access to a commercial frontier model — Anthropic's Fable 5 and

Mythos 5 — over a demonstrated safety-classifier jailbreak and a classified red-team finding about autonomous penetration capability [14]. CSA's own analysis of that episode, "The Fable 5 / Mythos 5 Export-Control Action," documents how a licensing directive originally scoped to address a specific safety failure produced a global service disruption because the provider had no mechanism to distinguish users by citizenship in real time, and how the incident established a new precedent: that the U.S. government is prepared to assert direct, model-specific authority over frontier AI access when it judges national security interests to be at stake [14]. Read together, the Fable-Mythos precedent and the distillation advisory describe two edges of the same emerging policy posture. One edge restricts who may access a U.S. frontier model. The other now asserts a government interest in how, and how much, foreign actors may learn from that access even when it is nominally authorized. Enterprises and AI providers should expect this posture to continue expanding rather than to represent a single, contained episode, and should plan governance and compliance programs around the expectation of further government intervention in frontier AI availability rather than around either event in isolation.

Contested Ground: The Legitimate-Use Defense

The advisory does not describe an uncontested set of facts, and treating it as such would misstate both the state of the evidence and the diplomatic stakes involved. China's Ministry of Commerce responded within twenty-four hours of the advisory's publication, rejecting the allegations as "groundless" and describing distillation as "a neutral technical means used by global model companies including U.S. firms," used to "enhance learning efficiency" and achieve "more efficient utilization of human knowledge" — language that deliberately mirrors the legitimate research and engineering use cases the technique genuinely serves [11]. The ministry further stated that Beijing would pursue "resolute countermeasures" if the United States uses the distillation allegations as a pretext to impose new restrictions on Chinese AI companies, without specifying what form those countermeasures might take [11]. This exchange occurred ahead of planned trade talks between the U.S. and Chinese heads of state, placing the advisory squarely inside an active and higher-stakes bilateral negotiation rather than in a purely technical or law-enforcement context [11].

The core of China's rebuttal — that distillation is a normal, widely used technique rather than inherently an act of theft — is not wrong as a general statement about the technology. U.S. frontier labs themselves use distillation routinely to produce smaller, cheaper model variants, and the broader open-source AI ecosystem depends on techniques that overlap substantially with what the advisory describes. What the advisory and the preceding corporate disclosures argue is not that distillation itself is illegitimate, but that the specific pattern observed — industrial account fraud, systematic evasion of geographic and usage restrictions, deliberate extraction of internal reasoning traces via jailbreak techniques, and operational infrastructure purpose-built to defeat detection — exceeds what any legitimate research or

commercial use of an authorized API account would require [1][8]. That is a defensible distinction in principle, but it is also one the public version of the advisory does not fully substantiate with auditable evidence, since the specific account-level and traffic-level data underlying the "billions of tokens" claim remains in the possession of the providers and the agencies rather than published for independent review [1].

Readers of this paper — particularly enterprise risk and compliance leaders — should treat the underlying factual dispute as genuinely unresolved at the level of precise scale and intent, even while treating the pattern of access obfuscation, account fraud, and evasion technique as well-corroborated across multiple independent sources spanning three major AI labs and three federal agencies [1][8][10]. The practical risk to enterprise security teams' defensive posture does not depend on resolving that dispute, though the dispute's ultimate resolution will materially affect export-control exposure, entity-list risk, and legal liability for the companies involved. Whether the extraction is state-directed espionage or aggressive-but-legal competitive practice, the operational reality — proxy-routed traffic at industrial scale defeating standard API controls — is the same, and it is the reality enterprise security teams and AI providers must build defenses against regardless of how the geopolitical dispute is ultimately adjudicated.

Implications for AI Providers and Enterprises

The advisory's recommended mitigations describe a defensive posture that, in our assessment, few organizations — including most frontier AI providers — had fully operationalized as recently as early 2026. Providers are asked to move beyond simple rate limiting toward behavioral analytics capable of distinguishing distillation activity from legitimate high-volume use — tracking subscription-to-usage ratios, detecting accounts that reach maximum usage immediately upon creation, and identifying the round-the-clock, low-variance usage patterns characteristic of automated extraction rather than human interaction [1]. They are further asked to build the capability to silently and selectively degrade output quality — reduced reasoning depth, alternative reasoning pathways, or calibrated output noise consistent with the differential-privacy approaches NIST recommends — for accounts suspected of distillation, without alerting those accounts to the fact that their service has been altered [1][17]. And they are asked to participate in cross-organization intelligence sharing sufficient to correlate distributed campaigns that no single provider's telemetry would reveal on its own, of the kind already underway through the Frontier Model Forum since April 2026 [1][18].

For enterprises that consume frontier AI models rather than build them, the advisory's direct relevance is less obvious but still material. Organizations that rely on third-party API aggregators or resellers to access frontier models at volume should treat the advisory's description of aggregator-based

obfuscation as a signal to review those relationships: a legitimate enterprise customer routing high-volume traffic through an aggregator may now be more likely to trigger the same behavioral flags the advisory recommends providers deploy against genuine distillation activity, creating a new category of false-positive business risk. Enterprises with foreign-national workforces or offshore development teams accessing frontier models should also read this advisory alongside the Fable-Mythos export-control precedent: both developments point toward a regulatory environment in which frontier AI access is increasingly conditioned, monitored, and subject to sudden government-directed change, and business continuity planning for AI-dependent workflows should account for that volatility as a standing risk rather than a one-time event, a conclusion CSA's own dependency-risk research reaches in greater detail [14] [19].

Foundation model providers and AI labs more broadly should treat this advisory as confirmation that model outputs, not just model weights and infrastructure, belong inside their intellectual property threat model. CSA's prior research on foundation model IP theft characterized model weights as the highest-value, most concentrated target in an AI lab's environment, but explicitly identified API-based extraction and distillation as a lower-sophistication, higher-volume complement to weight theft — one that requires no infrastructure breach at all, only sustained, well-resourced abuse of an authorized access channel [13]. The joint advisory reinforces that assessment and adds operational detail — the ATLAS mappings, the specific evasion techniques, the recommended detection indicators — that AI labs can use directly to extend existing IP-protection and insider-threat programs to cover output-level extraction, a threat category that conventional data-loss-prevention and access-control tooling was not designed to address.

CSA Resource Alignment

This advisory sits at the intersection of two bodies of CSA research that together provide the most directly relevant frameworks for organizations responding to it. CSA's "Foundation Model IP Theft: Threat Model for AI Labs" identifies foundation model weights, training data, and outputs as a new and concentrated class of intellectual property targeted by nation-state and competitive actors alike, and it specifically catalogs API-based distillation as a distinct, low-sophistication attack vector alongside infrastructure compromise and insider threat, recommending behavioral analytics for distillation detection rather than rate limiting alone — precisely the posture the September 2026 advisory now recommends government-wide [13]. Organizations building or extending an AI lab threat model in light of this advisory should treat that CSA research note as the direct technical companion to the government advisory: where the advisory tells defenders what a distillation campaign against their organization is likely to look like, CSA's threat model explains how to fold that detection capability into a broader model-provider security program aligned with the AI Controls Matrix's Model Provider domain [13].

CSA's analysis of "The Fable 5 / Mythos 5 Export-Control Action" provides the necessary governance and policy context for reading this advisory as part of a pattern rather than an isolated event. That research documents the first use of U.S. export control authority to suspend a deployed commercial frontier model over national-security concerns, and it establishes the precedent that frontier AI providers and their enterprise customers must now plan around direct, model-specific government intervention as a standing operational risk rather than a hypothetical one [14]. Enterprises assessing their exposure to this distillation advisory should read it together with that precedent, since both reflect the same underlying policy trajectory: increasing direct U.S. government involvement in who may access frontier AI models, on what terms, and with what downstream use. CSA's companion analysis, "Sovereign AI Risk: When Your AI Vendor Gets Export-Controlled," extends that precedent into a broader framework for enterprise dependency risk, arguing that model-agnostic routing and multi-provider strategies are the appropriate hedge against exactly the kind of sudden access interruption this advisory's enforcement posture could eventually produce [19].

Finally, the AI Controls Matrix (AICM) v1.1 remains the applicable governance baseline for operationalizing both the advisory's technical recommendations and the broader supply-chain and IP-protection concerns this paper raises. AICM's Model Provider domain addresses the access control, telemetry, and supply-chain integrity controls that behavioral distillation detection depends on, while its broader governance, risk, and compliance domains provide the structure enterprises need to evaluate AI vendor risk, including the export-control and geopolitical dependency risk this advisory and the Fable-Mythos precedent both illustrate [15]. Organizations without a mature AICM implementation should treat this advisory as a concrete forcing function: the detection, response, and information-sharing capabilities it recommends are not achievable without the underlying access-logging, identity, and monitoring controls AICM already specifies.

Conclusions and Recommendations

The joint NSA-CISA-FBI advisory on China-based AI distillation campaigns represents a formal government ratification of a threat pattern the AI industry itself had already spent roughly seven months documenting, escalating, and beginning to defend against. Its principal contribution is not new factual discovery but categorical reframing: by issuing the assessment as a joint interagency advisory structured around an established adversarial machine learning threat framework, the U.S. government has moved industrial-scale model distillation from the category of commercial dispute into the category of national security concern, with all the institutional, diplomatic, and regulatory consequences that shift implies. That reframing is contested — China's government has rejected it directly and threatened retaliation —

and the public evidence supporting the advisory's precise scale claims remains, appropriately, less transparent than the well-corroborated pattern of access obfuscation and evasion technique underlying it.

For AI providers, the advisory should accelerate investment in behavioral detection, selective output degradation, and cross-organization threat-intelligence sharing as standing operational disciplines rather than incident-specific responses. For enterprises, it should prompt a review of any AI access arrangements that route through third-party aggregators or resellers, and it should be read alongside the Fable-Mythos export-control precedent as part of a single, continuing trend toward direct government involvement in frontier AI availability. For policymakers, the unresolved tension at the center of this advisory — that the same technique the U.S. government now treats as a national security threat is also legitimate machine learning practice the industry it is trying to protect uses routinely — will need clearer legal and technical definition before enforcement mechanisms like export controls or entity-list designations can be applied to distillation activity without significant collateral effect on legitimate research and commercial practice. Organizations that wait for that definitional clarity before building detection and governance capability will find themselves behind an adversary the government has now formally assessed to be operating at industrial scale.

References

[1] Cybersecurity and Infrastructure Security Agency, National Security Agency, and Federal Bureau of Investigation. "China-Based Artificial Intelligence Companies Conducting Industrial-Scale Distillation Campaigns Against U.S. AI Companies (AA26-251A)." CISA, September 8, 2026.

[2] National Security Agency. "NSA and Others Warn China-Based AI Companies Are Distilling US Frontier AI Models." NSA Press Release, September 8, 2026.

[3] Abrams, Lawrence. "US Says Chinese Firms Extracted Billions of Tokens From Frontier AI Models." BleepingComputer, September 2026.

[4] SecurityWeek. "US Agencies Warn China Is Systematically Extracting Frontier AI Capabilities." SecurityWeek, September 2026.

[5] The Hacker News. "U.S. Agencies Accuse China AI Firms of Distilling Claude, GPT, Gemini, and Grok." The Hacker News, September 2026.

[6] Defense One. "China Is Trying to Steal US AI Models' Secrets, Intel Agencies Warn." Defense One, September 2026.

[7] Unite.AI. "NSA, CISA, FBI Warn China-Based AI Firms Distill US Frontier Models." Unite.AI, September 2026.

[8] CNBC. "Anthropic Accuses DeepSeek, Moonshot and MiniMax of Distillation Attacks on Claude." CNBC, February 24, 2026.

[9] CNBC. "Anthropic's Distillation Battle Turns to the Dark Web as China Concerns Swell." CNBC, September 3, 2026.

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

[11] Xinhua. "China Firmly Opposes U.S. Groundless Allegations on Chinese AI Firms: Commerce Ministry." Xinhua, September 10, 2026.

[12] Center for Strategic and International Studies. "DeepSeek, Huawei, Export Controls, and the Future of the U.S.-China AI Race." CSIS, 2026.

[13] Cloud Security Alliance AI Safety Initiative. "Foundation Model IP Theft: Threat Model for AI Labs." CSA Labs, May 17, 2026.

[14] Cloud Security Alliance AI Safety Initiative. "The Fable 5 / Mythos 5 Export-Control Action." CSA Labs, 2026.

[15] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." CSA, June 22, 2026.

[16] MITRE. "MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems." MITRE Corporation.

[17] National Institute of Standards and Technology. "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (NIST AI 100-2e2025)." NIST, March 24, 2025.

[18] Kasanmascheff, Markus. "OpenAI, Anthropic, Google Team Up to Stop Chinese AI Model Distillation." Winbuzzer, April 8, 2026.

[19] Cloud Security Alliance AI Safety Initiative. "Sovereign AI Risk: When Your AI Vendor Gets Export-Controlled." CSA Labs, July 2, 2026.

[20] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network." arXiv, March 9, 2015.