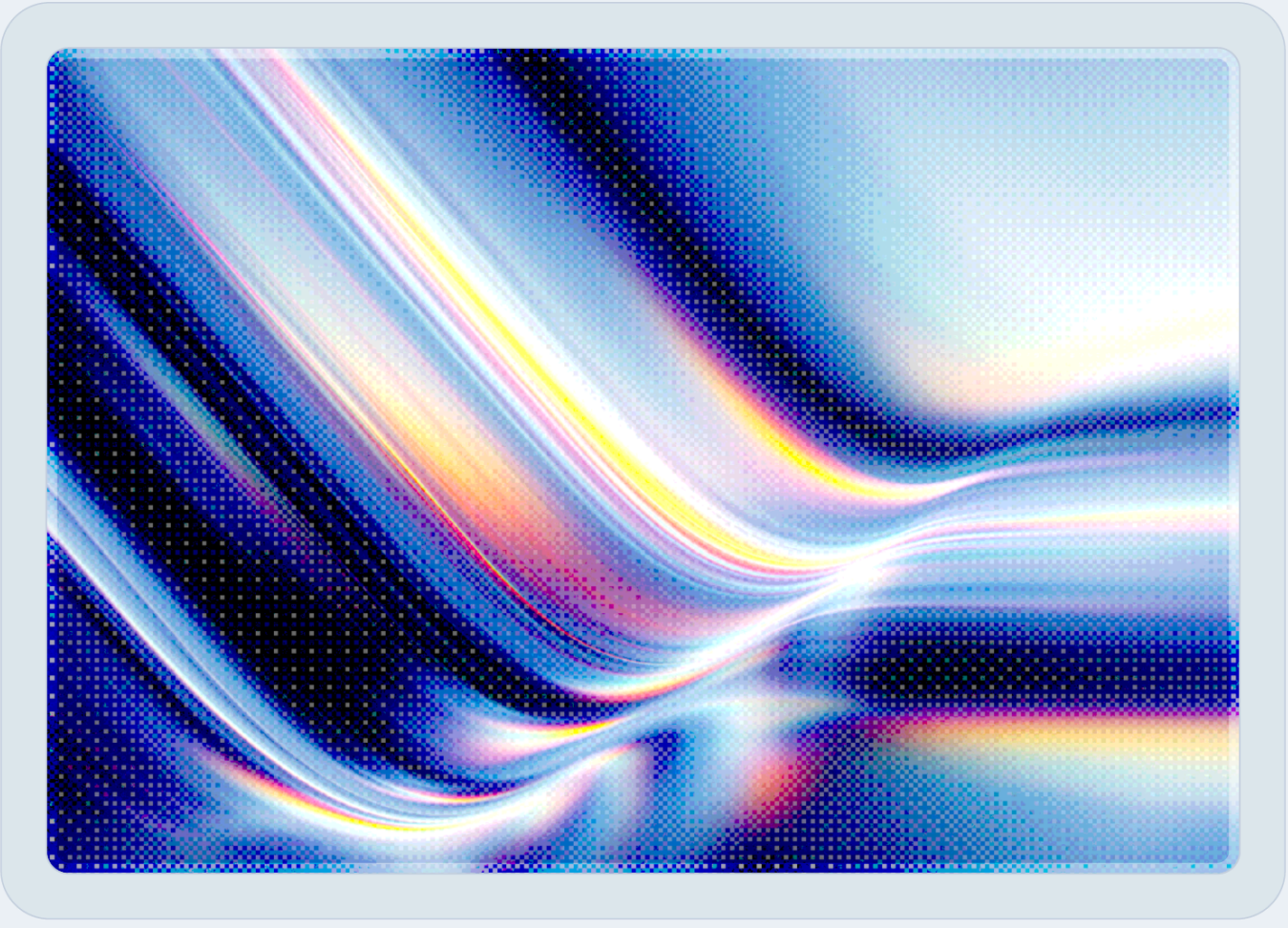
CSAI Foundation | Cloud Security Alliance

CISA Advisory: China's Industrial-Scale AI Distillation Campaign

Six Chinese AI Firms Named in Systematic Extraction of U.S. Frontier Models

2026-09-18

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

Key Takeaways

On September 8, 2026, the National Security Agency, the Cybersecurity and Infrastructure Security Agency, and the Federal Bureau of Investigation released joint advisory AA26-251A, warning that six China-based AI companies, DeepSeek, Moonshot AI, Alibaba, MiniMax, StepFun, and Z.AI, have run industrial-scale knowledge-distillation campaigns against U.S. frontier AI models since at least late 2024 [1][2]. The advisory describes distillation, a legitimate machine-learning technique for training smaller models to mimic a larger one's outputs, as having been weaponized through fraudulent account creation, jailbreak prompts designed to extract chain-of-thought reasoning, and "transfer station" API proxies that evade geographic access restrictions, together generating billions of tokens across millions of exchanges against variants of Claude, GPT, Gemini, and Grok [1][3]. The agencies assess that this activity forms the critical core, not merely a supplement, of the named companies' AI development strategy, and that it is proceeding with likely awareness of the Chinese government, violates the terms of service of the targeted U.S. providers, and materially strengthens capabilities with military and offensive cyber applications [1][3]. The advisory documents specific cases, including Alibaba distilling Claude-4, Claude Opus, Claude Sonnet, and GPT-5 in late 2025 to improve its Qwen model family's software engineering and customer-service dialogue, and StepFun distilling Claude Opus 4.1, Claude Opus 4.5, Claude Sonnet 4.5, Claude Haiku 4.5, and several GPT-5 variants between late 2025 and early 2026 to enhance its Step 4 model's coding and agentic functions [3]. The advisory also directly challenges DeepSeek's widely cited $5.6 million training-cost claim, stating it excludes the cost of capabilities acquired through distillation, and recommends that AI providers deploy behavioral detection of anomalous account and usage patterns, differential privacy and response-variation techniques, and cross-organization threat-intelligence sharing [1][4]. Security and AI governance teams, whether at model providers or at enterprises that consume frontier AI APIs, should treat this advisory as confirmation that query-based model extraction has moved from a theoretical threat vector to a documented, state-adjacent, industrial-scale practice with direct implications for API governance, vendor risk management, and export-control policy.

Background

Knowledge distillation is an established technique in machine learning research in which a smaller "student" model is trained to reproduce the input-output behavior of a larger "teacher" model, typically by training on the teacher's outputs rather than on the original data or architecture that produced them [1]. Used legitimately, distillation compresses expensive frontier capabilities into smaller, cheaper models for deployment on constrained hardware, and it does not require access to a model's weights or training data, only sustained access to its outputs through a public or authenticated interface such as an API. This is precisely what makes it attractive as an extraction vector: a well-resourced actor can approximate years of another organization's research and hundreds of millions to billions of dollars of compute spending by systematically querying a deployed model and training on the responses, without ever breaching the provider's infrastructure [1][8].

The joint advisory describes this dynamic playing out at scale against the leading U.S. frontier model families. Query campaigns extracted specific functional capabilities rather than general-purpose imitation: legal-domain reasoning, chain-of-thought and agentic task performance, software engineering and code generation, and creative or rule-following behaviors tuned through reinforcement learning and supervised fine-tuning were each targeted as discrete extraction objectives [1][3]. The operational tradecraft documented in the advisory goes well beyond simply issuing high volumes of API calls. Named companies allegedly used fraudulent accounts with sanitized or obfuscated metadata to obscure their origin, routed traffic through centralized "transfer station" proxies to bypass geographic and export-related access restrictions, exploited bulk commercial subscriptions to optimize per-token cost, and built automated failover systems that would redirect extraction traffic to alternate access pathways when one channel was blocked or rate-limited [1][5]. MiniMax was specifically observed redirecting its distillation infrastructure toward newly released Claude models within 24 hours of their public availability, indicating a standing operational capability rather than an ad hoc research effort [5]. The advisory frames the aggregate campaign duration as running from late 2024 through at least mid-2026, with query volumes described as ranging from thousands to millions of requests per targeted capability domain [1][4].

CISA Acting Director Nick Andersen framed the response directly: "We strongly urge AI companies to take immediate steps to safeguard their platforms against knowledge distillation campaigns" [4]. The three agencies were explicit that they view the scale and coordination of the activity as evidence of more than isolated commercial opportunism. Because the advisory states the campaigns are proceeding with likely Chinese government awareness and describes the resulting capability gains as relevant to military and cyberattack capacity, it arguably situates commercial AI distillation within the same national-security framing that has previously applied to export controls on model weights and AI-relevant semiconductor hardware, such as the Bureau of Industry and Security's ECCN 4E091 classification for certain trained

model weights [8]. In CSA's view, the advisory's core novelty is not that model extraction is possible — security researchers and CSA's own May 2026 analysis [8] had already described API-based distillation as a known, low-sophistication attack path — but that three U.S. national security agencies have now attributed a specific, sustained, and commercially significant campaign to six named companies with government-level visibility implied.

Security Analysis

In CSA's assessment, the advisory's most consequential claim for security teams is the assertion that distillation forms the critical core of the named companies' development strategy rather than a supplementary shortcut. That framing matters because it changes the threat model from opportunistic, bounded extraction attempts to a standing, well-funded, continuously operated capability with dedicated infrastructure, the transfer-station proxies, automated failover, and account-pool management described in the advisory are not artifacts of a single research project but components of persistent extraction tooling [1][5]. This validates a threat vector CSA's own May 2026 threat model for foundation-model intellectual-property theft identified as the lowest-sophistication and hardest-to-attribute path to model IP loss: unlike weight exfiltration or ML-toolchain compromise, API-based distillation requires no infrastructure breach at all, only sustained, well-disguised access to a public-facing endpoint the provider intentionally exposes [8]. The economics reinforce why this vector is difficult to deter through cost alone. CSA's prior analysis estimated frontier model training compute costs in the hundreds of millions to billions of dollars, and the advisory's direct challenge to DeepSeek's $5.6 million cost figure illustrates the same asymmetry in a live case: distillation lets a well-resourced adversary approximate years of proprietary research and reinforcement-learning investment for a fraction of the cost, provided sustained API access can be maintained [1][8].

The detection challenge described in the advisory is fundamentally one of distinguishing malicious extraction from legitimate high-volume usage, and the recommended indicators reflect that difficulty. Rather than relying on simple rate limits, which sophisticated actors can defeat by distributing load across proxies and account pools, the advisory recommends behavioral signals: subscription-to-usage ratios that reveal accounts consuming far more capacity than their billing tier would normally support, new accounts that immediately begin operating at maximum throughput rather than showing the gradual usage ramp typical of a new customer, and continuous 24/7 query patterns inconsistent with human-paced interaction [1][4][5]. This mirrors the behavioral-analytics approach CSA's foundation-model threat model recommended as a defense against API distillation specifically because rate limiting alone was assessed as insufficient [8]. The advisory's account of MiniMax redirecting extraction traffic to new

Claude releases within 24 hours also suggests that defenders should treat post-release monitoring windows as a heightened-risk period requiring tighter anomaly-detection thresholds rather than static, always-on limits calibrated for steady-state traffic.

One recommendation in the advisory deserves particular scrutiny from security and legal teams: serving deliberately "downgraded" model responses, or otherwise subtly altering outputs, to accounts suspected of distillation activity, without disclosing that alteration to the user [1][5]. This is a defensible technical countermeasure, degrading the quality of extracted training data reduces the value of a successful distillation attempt, but it is also a form of covert, undisclosed service degradation that sits in tension with transparency norms many providers otherwise commit to in their terms of service and public trust communications. Organizations implementing this mitigation should weigh it against contractual and disclosure obligations to legitimate customers who might be misidentified by behavioral heuristics, since false positives in this kind of detection carry a direct product-quality cost that is, by design, invisible to the affected customer. More broadly, the advisory's emphasis on cross-organization intelligence sharing, correlating account infrastructure, proxy IP ranges, and third-party billing services across competing model providers, implies the agencies see a coordination gap among providers. An AI-specific information-sharing and analysis center has reportedly been in development under CISA since 2025, but had no confirmed public launch as of early 2026, and CSA has not independently verified the current maturity of any such mechanism; organizations should treat the existence of a standing, cross-provider channel for this kind of behavioral indicator data as an open question rather than a settled gap, and in the meantime should assume each provider is largely defending against a shared, coordinated adversary in isolation [1][4].

Recommendations

Immediate Actions

AI model providers should audit existing API telemetry for the specific behavioral indicators named in the advisory: subscription-to-usage ratio anomalies, new accounts reaching maximum throughput immediately after signup, and sustained 24/7 query patterns inconsistent with typical customer behavior, and should establish alerting thresholds calibrated to these patterns rather than relying solely on static rate limits [1][4]. Providers should also review account-verification and enterprise-billing processes for the specific evasion patterns described in the advisory, particularly bulk subscription purchasing and use of third-party API aggregators or "transfer station" proxies, since these were identified as primary mechanisms for evading existing geographic and access controls [1][5]. Enterprises that consume frontier model APIs at scale, particularly those with usage patterns that might resemble

the flagged behavioral indicators, such as automated pipelines running continuously or bulk-purchased enterprise seats, should proactively engage their model providers to ensure legitimate high-volume usage is not inadvertently throttled or degraded by these new anomaly-detection measures.

Short-Term Mitigations

Security and product teams at model-provider organizations should evaluate the specific technical countermeasures the advisory references from MITRE ATLAS and NIST's adversarial machine learning taxonomy, including query rate limiting tied to behavioral risk scoring rather than flat thresholds, output obfuscation for chain-of-thought and reasoning traces that are disproportionately valuable to distillation, and differential-privacy techniques applied to responses served to higher-risk accounts [1][6][7]. Any organization considering the advisory's response-alteration recommendation, serving modified or "downgraded" outputs to suspected extraction accounts, should route that decision through legal and trust-and-safety review given the transparency trade-offs involved, and should build in an appeals or override path for legitimate customers who trigger the heuristics. Providers should also begin establishing, or joining, a cross-organization channel for sharing infrastructure indicators, such as proxy IP ranges, known transfer-station domains, and third-party billing services associated with extraction activity, since the advisory's core finding is that this activity is distributed across multiple providers simultaneously and no single provider has full visibility into the pattern.

Strategic Considerations

Enterprises that rely on frontier AI providers for competitive differentiation should factor distillation-driven capability convergence into vendor risk and competitive-strategy assessments: if competitor capabilities can be approximated through extraction rather than independent development, the assumption that a proprietary model API provides a durable, exclusive advantage requires re-examination, particularly for capabilities exposed through less-restricted API tiers. Export-control and trade-policy teams should track whether the scope of controls like ECCN 4E091, currently focused on model weights [8], is extended to address API-based extraction, since the advisory demonstrates that a determined actor can approximate a substantial share of a frontier model's value without ever obtaining the underlying weights [1]. Finally, organizations evaluating open-weight models originating from the companies named in the advisory should treat the provenance and IP pedigree of those models as an active due-diligence question rather than an assumption, given that models incorporating extracted capabilities from U.S. frontier systems may carry undisclosed legal, licensing, or supply-chain risk when integrated into downstream enterprise products.

CSA Resource Alignment

This advisory validates and extends CSA's own Foundation Model IP Theft: Threat Model for AI Labs, which identified API-based distillation as a low-sophistication, high-impact extraction path against foundation models and recommended behavioral-analytics detection over rate limiting alone, a recommendation the CISA/NSA/FBI advisory now echoes with specific, attributed campaign data [8]. Organizations building or refining an AI-lab threat model should treat the two documents as complementary: CSA's paper provides the broader threat-model structure spanning model weights, ML infrastructure, and supply chain, while the joint government advisory supplies concrete, named evidence of the distillation vector operating at nation-state-adjacent scale. Because the named companies' activity implicates AI model providers' security and governance obligations directly, CSA's AI Controls Matrix (AICM) v1.1 [9], and specifically its Model Provider (MP) domain, gives security and compliance teams a structured basis for evaluating whether their own or their vendors' API governance, abuse-detection, and access-control practices address the extraction techniques the advisory documents, rather than treating distillation defense as a purely ad hoc engineering decision. CSA's AICMv1.1 Auditing Guidelines for Model Providers (MP) [10] operationalizes this MP-domain alignment further, providing control-by-control auditing guidance that maps directly onto the abuse-detection and access-control practices the advisory recommends. Enterprises assessing exposure to distilled or extraction-derived AI capabilities in their supply chain should use these resources together when conducting AI vendor risk reviews, particularly for vendors whose model lineage or training provenance is not independently verifiable.

References

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

[2] CISA. "CISA, NSA and FBI Warn of China-Based AI Companies Targeting US AI Models with Industrial-Scale Knowledge Distillation Campaigns to Shortcut AI Development." Cybersecurity and Infrastructure Security Agency, September 8, 2026.

[3] National Security Agency. "NSA and Others Warn China-Based AI Companies are Distilling U.S. Frontier AI Models." National Security Agency, September 8, 2026.

[4] Help Net Security. "Chinese AI firms are siphoning capabilities from American models, CISA warns." Help Net Security, September 9, 2026.

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

[6] MITRE. "MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems." MITRE Corporation, accessed September 18, 2026.

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

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

[9] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2026.

[10] Cloud Security Alliance. "AICMv1.1 Auditing Guidelines for Model Providers (MP)." Cloud Security Alliance, 2026.