



China-Based Artificial Intelligence Companies Conducting Industrial-Scale Distillation Campaigns Against U.S. AI Companies

Executive summary

China-based artificial intelligence (AI) companies are conducting systematic extraction of proprietary functionalities and capabilities of U.S. AI companies' models through industrial-scale knowledge distillation campaigns that form the core—not merely a supplement—of their AI development strategy. While “distillation” is recognized as a legitimate and useful technique in AI research, China-based AI companies are engaging in aggressive, malicious, and targeted distillation activities at an industrial scale that extract restricted proprietary functionalities and capabilities of U.S. frontier AI models. The National Security Agency (NSA), Cybersecurity and Infrastructure Security Agency (CISA), and Federal Bureau of Investigation (FBI) (hereafter referred to as the authoring agencies) are releasing this joint Cybersecurity Advisory to alert organizations about these malicious activities and techniques and recommend mitigations to reduce their potential impact.

Likely with Chinese government awareness, DeepSeek, Moonshot AI, Alibaba, MiniMax, StepFun, and Z.AI extracted billions of tokens across millions of exchanges/requests from U.S. frontier AI models, including variants of Claude, GPT, Gemini, and Grok, since at least late 2024. DeepSeek has conducted organized campaigns since at least 2024 targeting reasoning capabilities, specialized optimizations, and domain-specific functions to train its R1 and V3 models. Alibaba leveraged industrial-scale distillation to improve the company's Qwen family of AI models. Moonshot AI, MiniMax, Stepfun, and Z.AI also engaged in malicious knowledge distillation of U.S. AI companies' models.

China-based AI companies route distillation requests through multiple pathways to gain unauthorized access, consequently violating U.S. AI companies' terms of use. These pathways include native application programming interfaces (APIs), remote cloud

TLP: CLEAR

providers, and third-party aggregators that automatically obfuscate user metadata to avoid detection. Further, China-based AI companies use a gray market of proxies known as “transfer stations” to bypass U.S. AI companies’ geographic restrictions, breach terms of use, evade safeguards, and undermine traceability. China-based AI companies achieve cost savings for their industrial-scale distillation campaigns through bulk procurement of the U.S. AI companies’ premium subscriptions shared across teams of developers. Advanced industrial-scale distillation tactics include chain-of-thought (CoT) reasoning extraction, automated failover between pathways during blocking attempts, and sophisticated quality evaluation frameworks to detect defensive countermeasures. China-based AI companies that conduct industrial-scale distillation against U.S. AI models see significantly shorter AI development timelines and reduced financial expenditures in training a frontier model.

China-based AI companies deliberately distribute operations across multiple providers, platforms, and pathways to avoid single-point detection. They also attempt to distill the best capabilities and proprietary features of each U.S. frontier model to train their China-based AI models. This represents systematic extraction of proprietary functionalities and capabilities threatening U.S. technological leadership. Addressing industrial-scale distillation merits a coordinated response across the AI ecosystem, including effective information-sharing, spanning the U.S. Government, private industry, and allied nations.

The authoring agencies recommend U.S. AI companies take three immediate actions:

1. **Implement comprehensive detection and mitigation:** Detect anomalous and malicious prompts, accounts, networks, and behaviors. Additionally, monitor subscription-to-usage ratios, immediate maximum usage from new accounts, and enterprise-scale throughput patterns.
2. **Deploy targeted response changes:** Subtly alter responses for suspected malicious distillation attempts to attenuate the payoffs to companies conducting industrial-scale distillation campaigns.
3. **Establish cross-organization intelligence sharing:** Correlate activity across model providers, cloud platforms, and API aggregators to reveal distributed campaigns.

TLP: CLEAR

Attribution

Since at least late 2024, China-based AI companies, including DeepSeek (DeepSeek Artificial Intelligence Technology Research Co., Ltd.), Moonshot AI (Beijing Moonshot Technology Co., Ltd.), Alibaba Group, MiniMax (Shanghai MiniMax Co., Ltd.), StepFun (Shanghai Jieyue Xingchen Intelligence Technology Co., Ltd.), and Z.AI, have conducted high-volume knowledge distillation campaigns against several U.S. AI companies. The sheer scale of these campaigns and their sophistication indicate that distillation is not a supplement to these companies' AI model development, but the critical core of it.

Likely with the knowledge of the Chinese government, the China-based AI sector has turned to a comprehensive distillation strategy in an attempt to bridge the technological and performance gaps between their AI models and U.S. frontier AI models. To access U.S. AI companies' application programming interfaces (APIs), China-based AI companies use a gray market of API proxies known as "transfer stations" to bypass U.S. AI companies' regional restrictions, breach terms of use, evade safeguards, and undermine traceability.

DeepSeek

DeepSeek has been conducting an organized distillation campaign against U.S. AI companies' frontier AI models since at least late 2024 to generate synthetic training data for its models, including R1, released in early 2025. The company targeted specific knowledge domains to extract proprietary functionality and reasoning capabilities to reduce their compute and research costs. DeepSeek's publicly quoted training costs of \$5.6M are misleading as it does not include the true cost of the data acquired through extensive malicious distillation.¹

Between late 2024 and mid-2025, DeepSeek distilled specialized training data and capabilities from the following U.S. frontier AI company models to train their R1 and V3 models:

- Claude 3.7
- Claude Sonnet 4
- Claude Sonnet 4.5

¹ Publicly quoted training costs are from "[DeepSeek-V3 Technical Report](#)"

TLP: CLEAR

- Claude Opus 4.1
- Gemini 2.5 Pro Preview
- Gemini 2.5 Flash Preview
- GPT-4
- GPT-4o
- GPT-4 Mini
- GPT-4 Nano
- GPT-5
- Grok 4

The specific knowledge and capabilities distilled included:

- Legal specialization optimization
- API rule-driven tasks
- Writing using CoT drafts
- Agentic functions
- Question and answer optimization
- Coach/assistant capabilities
- Functional creation optimization
- Supervised fine-tuning (SFT) optimization
- Creative and occupational writing optimization

Moonshot AI

Moonshot AI has conducted a widespread distillation campaign against U.S. frontier AI companies since at least mid-2025. Notably, Moonshot AI extracted significant Claude Fable 5 data to train its Kimi-K3 model and GPT-4o data to train its Kimi-K2 model. The company has used the following models to distill SFT optimization, reinforcement learning (RL), software engineering, and math capabilities:

- Claude Opus 4.1
- Claude Sonnet 3.7
- Claude Sonnet 4
- Claude Sonnet 4.5
- Claude Sonnet 4.5 Thinking
- Claude Fable 5
- GPT-oss-20b
- GPT-3

TLP: CLEAR

TLP:CLEAR

- GPT-4o
- GPT-4o mini
- GPT-5
- GPT-5 Codex
- GPT-5 Pro
- Gemini 2.5 Flash
- Gemini 2.5 Flash-Image
- Gemini 2.5 Pro
- Nano Banana
- Grok Code Fast-1

Other companies

Several other China-based AI companies, including Alibaba, MiniMax, StepFun, and Z.AI have also leveraged distillation techniques to build their AI models. In late 2025, Alibaba distilled Claude-4, Claude Opus, Claude Sonnet, and GPT-5 to improve their AI models' software engineering skills, customer service dialogue functionality, image/character creation, and integration of RL, SFT, and distillation capabilities.

In late 2025, MiniMax distilled CoT reasoning, RL, SFT, and software engineering capabilities to improve its M2 model from Claude Code, Claude Sonnet 4, Claude Opus, Gemini 1, Gemini 2.5 Pro, and Gemini 3 Pro. MiniMax used Claude Code for internal software development tasks, including code generation, analysis, and refinement. MiniMax even used prompt injections to try to trick Claude Code into believing it was a MiniMax product.

Between late 2025 and early 2026, StepFun distilled data from Claude Opus 4.1 and 4.5, Claude Sonnet 4.5, Claude Haiku 4.5, GPT-5 Mini, GPT-5 Pro, GPT-5.1, GPT-5.1 Codex, and GPT-5.2 to improve its Step 4 model's coding and agentic functions. By mid-2026, Z.AI had distilled billions of tokens of GPT-5.5 data and Claude Opus 4.8 data to develop the CoT reasoning capabilities of its model.

TLP:CLEAR

Table 1: China-based AI Companies Engaged in Knowledge Distillation Against U.S. AI Companies

(From at least 2024-2026)

China-based AI Company	U.S. AI Models Distilled	Functionalities and Domains Distilled
DeepSeek (DeepSeek Artificial Intelligence Technology Research Co., Ltd.) 深度求索AI基础技术研究有限公司	• Claude Sonnet 3.7 • Claude Sonnet 4 • Claude Sonnet 4.5 • Claude Opus 4.1 • Gemini 2 • Gemini 2.5 Pro Preview • Gemini 2.5 Flash Preview • GPT-4 • GPT-4o • GPT-4 Mini • GPT-4 Nano • GPT-5 • Grok 3 Mini • Grok 4	• Legal specialization optimization • API rule-driven tasks • Writing using CoT drafts • Question and answer optimization • Coach/assistant capabilities • Functional creation optimization • SFT optimization • Agentic capabilities • Creative and occupational writing optimization
Moonshot AI (Beijing Moonshot Technology Co., Ltd.) 北京揽月星辰科技有限公司	• Claude Opus 4.1 • Claude Sonnet 3.7 • Claude Sonnet 4 • Claude Sonnet 4.5 • Claude Sonnet 4.5 Thinking • Claude Fable 5 • GPT-oss-20b; • GPT-3 • GPT-4o mini • GPT-5 • GPT-5 Codex • GPT-5 Pro • Gemini 2.5 Flash • Gemini 2.5 Flash-Image • Gemini 2.5 Pro • Nano Banana • xAI Grok Code Fast-1	• SFT • RL • Software engineering • Math capabilities

TLP:CLEAR

TLP: CLEAR

Alibaba 阿里集团	• Claude 4 • Claude Sonnet • GPT-5	• Customer service dialogue • Virtual character creation • SFT, RL, and distillation training • Evaluating and training datasets • End-to-end agentic workflows • Software engineering
MiniMax (Shanghai MiniMax Co., Ltd.) 上海稀宇极智科技有限公司	• Claude Code • Claude Sonnet 4 • Claude Opus 4.5 • Gemini 1 • Gemini 2.5 Pro • Gemini 3 Pro • GPT-5	• CoT reasoning • Agentic functionality • Code review • SFT dataset refinement • Software engineering tasks
StepFun (Shanghai Jieyue Xingchen Intelligence Technology Co., Ltd.) 上海阶跃星辰智能科技有限公司	• Claude Opus 4.1 • Claude Opus 4.5 • Claude Sonnet 4.5 • Claude Haiku 4.5 • GPT-5 Mini • GPT-5 Pro • GPT-5.1 • GPT-5.1 Codex • GPT-5.1 Codex Mini • GPT-5.2	• Code development • Agentic functions
Z.AI 北京智谱华章科技有限公司	• GPT-5.5 • Claude Opus 4.8	• CoT reasoning

TLP:CLEAR

Tactics, techniques, and procedures

China-based AI companies employ sophisticated tactics, techniques, and procedures (TTPs). These TTPs map to the [MITRE® ATLAS™](#)² framework, progressing through multiple adversary lifecycle phases from initial access through exfiltration. The China-based AI companies using these techniques include DeepSeek, Moonshot AI, MiniMax, StepFun, Z.AI, and other China-based AI companies targeting U.S. frontier AI models.

Table 2: MITRE ATLAS Mappings		
TTP Title	ID	Description
Resource Development		
Acquire Infrastructure	AML.T0008	China-based entities establish and maintain sophisticated infrastructure supporting sustained extraction operations through tiered budget management and diverse supplier relationships. China-based entities circumvent both Chinese and U.S. AI access controls through a large gray market of API proxies, or “transfer stations,” which resell access to frontier models at a fraction of the official price. In doing so, they create a scalable mechanism for evading provider safeguards and eroding traceability.
AI Model Access		
AI Model Inference API Access	AML.T0040	China-based entities have been exploiting AI model inference APIs through the creation of fraudulent accounts that are not registered to legitimate users. These actors leverage multiple accounts with similar registration details and payment methods, frequently switch between various AI models, and utilize third-party API aggregator services. Additionally, they execute

² MITRE is a registered trademark of The MITRE Corporation. MITRE ATLAS is a trademark of The MITRE Corporation.

TLP:CLEAR

TLP:CLEAR

		highly coordinated queries featuring identical or similar prompt texts, demonstrating a sophistication indicative of advanced AI research. The sheer volume of requests, ranging from thousands to millions on similar topics, far exceeds legitimate use, raising significant concerns about potential misuse and compromising the integrity of AI systems.
Execution / Privilege Escalation / Defense Evasion		
LLM Prompt Injection	AML.T0051	China-based entities have conducted prompt injection techniques against large language models (LLMs) by inserting prompts specifically designed for jailbreaking.
LLM Jailbreak	AML.T0054	China-based entities craft prompts forcing models to reveal their hidden CoT reasoning (CoT or step-by-step internal reasoning that enables greater capabilities) despite U.S. models restricting CoT output visibility to users. DeepSeek employed prompts instructing models to imagine and articulate the internal reasoning behind completed responses and write it out step by step. This CoT data teaches student models, not just factual knowledge, but reasoning methodologies for complex agentic tasks, coding challenges, and logical proofs.
Discovery		
Discovery	AML.TA0008	China-based entities employ aggressive, adaptive discovery to systematically identify valuable extractable data. China-based entities demonstrate rapid operational adaptation. MiniMax redirected exchanges to a new Claude model within 24 hours of release, demonstrating real-time

TLP:CLEAR

TLP:CLEAR

		provider monitoring and pre-positioned infrastructure for immediate retargeting.
AI Attack Staging		
Verify Attack	AML.T0042	China-based entities deploy production-grade automated quality assurance pipelines with multi-modal validation, enabling rapid detection of degraded outputs and differentiation of service issues from defensive data degradation.
Collection		
Collection	AML.TA0009	China-based entities systematically collect outputs to generate synthetic training datasets through continuous API querying, targeting specific knowledge domains rather than indiscriminate gathering. Moonshot AI used millions of exchanges targeting agentic reasoning/tool use, coding/data analysis, computer-use agent development, and computer vision, evolving from text-based distillation to extracting logical frameworks, enabling tool interaction and visual processing. DeepSeek used queries targeting reasoning capabilities, rubric-based grading tasks (reward model function), and censorship-safe query rewriting, extracting how U.S. models evaluate response quality. Campaigns span days to months with query volumes in the thousands to millions per domain, far exceeding legitimate research or development use cases.
Exfiltration		
Exfiltration via AI Inference API:	AML.T0024.002	China-based entities have been collecting U.S. frontier LLMs' inferences into datasets, which can be used to train their models to mimic the

TLP:CLEAR

TLP:CLEAR

Extract AI Model		behavior and performance of these LLMs.
Impact		
External Harms	AML.T0048	China-based entities inflict financial harm through systematic extraction of proprietary functionality and capabilities, causing significant economic losses. Extracting capabilities worth billions in development costs while undermining competitive advantages represents a strategic economic threat to fair technological competition and U.S. technological leadership.

Novel TTPs

China-based AI companies leverage techniques not in MITRE ATLAS, demonstrating significant organizational investment, operational maturity, and adaptive capability development distinguishing these campaigns from opportunistic exploitation.

Novel TTP 1: Regional restriction evasion and subscription exploitation

Some U.S. frontier AI models are restricted for use; however, China-based AI companies access U.S. frontier AI models by employing various means to bypass the regional restrictions.

After bypassing the restriction, China-based AI companies create user accounts obfuscating their country of origin and subsequently procure bulk premium AI subscription services.

StepFun structured access around pools of accounts with employees running multiple concurrent sessions, implementing load distribution to prevent quota depletion. Daily budget allocations per automated agent started at moderate levels, scaling significantly as operations matured.

Detection indicators include:

- shared accounts from multiple IPs/user agents,
- 24/7 sustained usage without human variation/idle periods,
- anomalous subscription-to-API usage ratios, and

TLP:CLEAR

TLP:CLEAR

- new subscriptions immediately at maximum usage as opposed to gradual AI adoption.

Novel TTP 2: Centralized request routing infrastructure

China-based AI companies deploy sophisticated tools that enable unified control and scalable implementation for evasion at scale. This provides model/provider abstraction, real-time health monitoring, centralized quota enforcement, and automated sanitization.

China-based AI companies manage routing systems to external AI models for distillation. These routing systems direct requests through multiple pathways: native APIs, cloud providers, third-party aggregators, third-party relays, and vendor account pools.

Detection indicators include:

- consistent operational patterns across diverse account pools and
- correlated timing/behavior across different pathways indicating unified orchestration.

Novel TTP 3: Automated request metadata sanitization

China-based AI companies implement automated sanitization to systematically remove organizational identifiers. This differs from [AML.T0065](#) (LLM Prompt Crafting) by operating at an infrastructure layer with automated enforcement instead of manual modification.

Detection indicators include:

- sudden behavioral changes following disclosures/sharing, especially abrupt disappearance of previously consistent metadata;
- absence of expected markers in high-volume campaigns where scale suggests institutional activity; and
- generic/randomized patterns replacing consistent organizational indicators.

Novel TTP 4: Systematic quota and cost optimization

China-based AI companies systematically minimize API costs through pathway selection prioritizing cost-efficiency, centralized quota allocation/budget alignment, and account segmentation by purpose.

TLP:CLEAR

TLP:CLEAR

Detection indicators include:

- new accounts with anomalously high immediate hit rates suggesting bulk deployment with pre-engineered templates,
- usage optimized for cache maximization versus task diversity, and
- coordinated pathway switching responding to pricing/rate changes indicating centralized decision-making.

Mitigations

Coordinated, ecosystem-wide responses extending beyond individual company measures can help address knowledge distillation campaigns. The mitigations below incorporate mitigations from the MITRE ATLAS and National Institute of Standards and Technology (NIST) AI frameworks. Collaboration across the broader AI ecosystem, including cloud providers, API aggregators, and infrastructure providers, can enable a coordinated defense against malicious knowledge distillation campaigns.

Behavioral detection and monitoring

China-based AI companies leverage premium subscriptions to U.S. frontier models for knowledge distillation campaigns and code development. U.S. companies should strengthen identity verification for accounts and track individual subscriptions with enterprise-scale throughput, accounts deviating from legitimate patterns, and new accounts immediately at maximum usage versus a gradual ramp-up or with consistent quota exhaustion.

Response alteration for suspected distillation activity

Employing targeted changes in response to high-confidence malicious distillation requests can impose meaningful costs on knowledge distillation campaigns. Response changes, such as including differential privacy or using less sophisticated “downgraded” models to respond to distillation requests, can help protect U.S. proprietary functionalities and capabilities and reduce payoffs from distillation attempts.

Implementation strategies

When suspecting a malicious distillation campaign, consider varying changes to responses across requests to complicate response quality evaluations, such that the

TLP:CLEAR

TLP:CLEAR

subtle changes avoid triggering obvious alerts. Reducing reasoning depth, presenting correct information with different reasoning, or stylistic inconsistencies may evade detection while reducing training usefulness.

Avoid informing China-based AI company users suspected of distillation campaigns of a switch to a downgraded model. Informing malicious distillers would enable them to improve their defense evasions and indicate when to roll back training. Instead, alter responses to users confirmed to be querying frontier models specifically for malicious knowledge distillation campaigns without informing them. In contrast, AI safety researchers and third-party evaluators should be informed of model changes while still applying strong distillation mitigations.

Cross-organization information sharing and ecosystem coordination

Sharing information about distillation campaigns, such as indicators of infrastructure distributing operations across multiple providers, platforms, and pathways, can improve individual companies' detection efforts. Industry disclosures document proxy networks managing tens of thousands of fraudulent accounts simultaneously, mixing distillation with unrelated customer requests across multiple providers. Community collaboration could provide defenders with more comprehensive visibility across the native APIs, cloud endpoints, and aggregators.

Sharing information about distillation enables and enhances correlation otherwise unachievable by individual organizations, through sharing infrastructure indicators (IPs, domains, third-party service providers) and behavioral indicators (timing correlations, query volume patterns).

Multi-source correlated activity enables more confident attribution of malicious knowledge distillation campaigns, justifying response degradation with lower-to-no legitimate user risk.

Sharing infrastructure and behavioral indicators between cloud providers, model aggregators, and model providers can make distributed infrastructure visible as coordinated campaigns versus isolated anomalies. Additionally, sharing can provide cloud and routing companies with actionable indicators for identifying and mitigating malicious activity.

TLP:CLEAR

TLP:CLEAR

MITRE ATLAS mitigations

- [AML.M0015](#) - Predictive AI Adversarial Input Detection: Detect/block atypical queries deviating from benign patterns, exhibiting previous adversary technique characteristics, or originating from malicious IPs.
- [AML.M0004](#) - Limit AI Service Query Volume and Rate: Per-key/IP quotas, rate limits, progressive throttling. Adversaries seem to be sensitive to rate limits since they implement sophisticated strategies to work within constraints.
- [AML.M0019](#) - Control Access to AI Models and Data in Production: User verification, authenticated API access, policy monitoring. This addresses fraudulent account pool exploitation.
- [AML.M0024](#) - AI Telemetry Logging: Log inputs/outputs for threat detection/forensics. This is foundational for behavioral detection and enables correlation with intelligence.
- [AML.M0002](#) - Predictive AI Output Obfuscation: Reduce fidelity of responses (withhold logits/confidences, shorten responses, targeted redaction). Balance security with user experience.
- [AML.M0035](#) – AI Red Team: Adversarial testing, extraction simulation, telemetry monitoring. Validates detection efficacy.
- [AML.M0015](#) - Predictive AI Adversarial Input Detection: Sanitize/validate inputs preventing prompt injections. This addresses jailbreak and injection attempts to elicit reasoning traces and system prompts.
- [AML.M0000](#) - Limit Public Information Release: Limit disclosure of architecture, prompt templates, and system instructions.
- [AML.M0001](#) - Limit Model Artifact Release: Limit release of data, algorithms, architectures, and model checkpoints.
- [AML.M0003](#) - Predictive AI Model Hardening: Use adversarial training and defensive distillation to increase jailbreak difficulty.
- [AML.M0006](#) - Predictive AI Ensembles: Use multiple models so extracting one yields a less usable clone.

TLP:CLEAR

TLP:CLEAR

NIST AI 100-2e2025: Adversarial machine learning mitigations

Mitigations in NIST's "[Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations](#)" (NIST AI 100-2e2025) also apply to malicious distillation, including differential privacy, pre- and post-training interventions, and prompt instruction/formatting.

Differential privacy

Differential Privacy (DP) provides mathematically rigorous protection against inference and distillation techniques by adding calibrated noise to model outputs and preventing malicious actors from extracting training data membership information and other sensitive model information, such as decision boundaries or signals that could help reconstruct private data. This protection is governed by privacy parameters that define a finite privacy budget, where each query consumes part of the model's available privacy protection and repeated querying steadily reduces the remaining privacy reserve.

As that budget is consumed through accumulated queries, the model must either add more noise to preserve privacy, restrict further queries, or accept reduced privacy protection. This creates a fundamental noise-versus-utility tradeoff, where stronger privacy protection requires more noise, which can lower prediction precision and business usefulness, while less noise improves utility but increases vulnerability to compromise techniques, such as membership inference, model extraction, or inversion.

In practice, the right balance requires careful tuning and empirical auditing, because theoretical privacy settings do not always predict real-world accuracy impact, particularly for complex models or high-dimensional outputs that require substantially more noise to achieve equivalent protection, or when facing adaptive actors. As a result, DP is often strengthened with complementary controls such as query rate limiting, response aggregation, and monitoring.

Pre/post-training interventions

A range of training strategies have been proposed to increase the difficulty of accessing harmful capabilities through prompt injection, including safety training during pre-training or post training, adversarial training methods, and other methods to make jailbreak techniques more difficult.

TLP:CLEAR

TLP:CLEAR

Prompt instruction/formatting

Model instructions can cue the model to treat user input carefully, such as by wrapping user input in XML tags, appending specific instructions to the prompt, or otherwise attempting to clearly separate instructions from user prompts to mitigate distillation and make prompt injection or jailbreaking less effective.

References

- [Anthropic: Detecting and preventing distillation attacks](#)
- [Google: GTIG AI Threat Tracker: Distillation, Experimentation, and \(Continued\) Integration of AI for Adversarial Use](#)
- [NIST AI 100-2e2025: Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations](#)
- [OpenAI: RE: Updated Stakes for American-Led, Democratic AI](#)
- [The Decoder: How China's gray market sells Claude tokens at a fraction of the price](#)
- [White House National Security Presidential Memorandum 11 \(NSPM-11\): Artificial Intelligence in the National Security Enterprise](#)
- [White House National Science and Technology Memorandum 4 \(NSTM-4\): Adversarial Distillation of American AI Models](#)
- [White House Office of Science and Technology Policy post on X](#)

Disclaimer of endorsement

The information and opinions contained in this document are provided "as is" and without any warranties or guarantees. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the United States Government, and this guidance shall not be used for advertising or product endorsement purposes.

Purpose

This document was developed in furtherance of the authoring agencies' cybersecurity missions, including their responsibilities to identify and disseminate threats and to develop and issue cybersecurity specifications and mitigations. This information may be shared broadly to reach all appropriate stakeholders.

Contact

- **National Security Agency**
Cybersecurity Report Feedback: CybersecurityReports@nsa.gov

TLP:CLEAR

Joint CSA | China-Based AI Companies Conducting Industrial-Scale Distillation Campaigns Against U.S. AI Companies

TLP:CLEAR

Defense Industrial Base Inquiries and Cybersecurity Services: DIB_Defense@cyber.nsa.gov
Media Inquiries / Press Desk: NSA Media Relations: 443-634-0721, MediaRelations@nsa.gov

- **Cybersecurity and Infrastructure Security Agency**
CISA's 24/7 Operations Center (contact@cisa.dhs.gov), or by calling 1-844-Say-CISA (1-844-729-2472).
- **Federal Bureau of Investigation**
If you or someone you know has fallen victim to this campaign, file a complaint with [IC3](#).

TLP:CLEAR