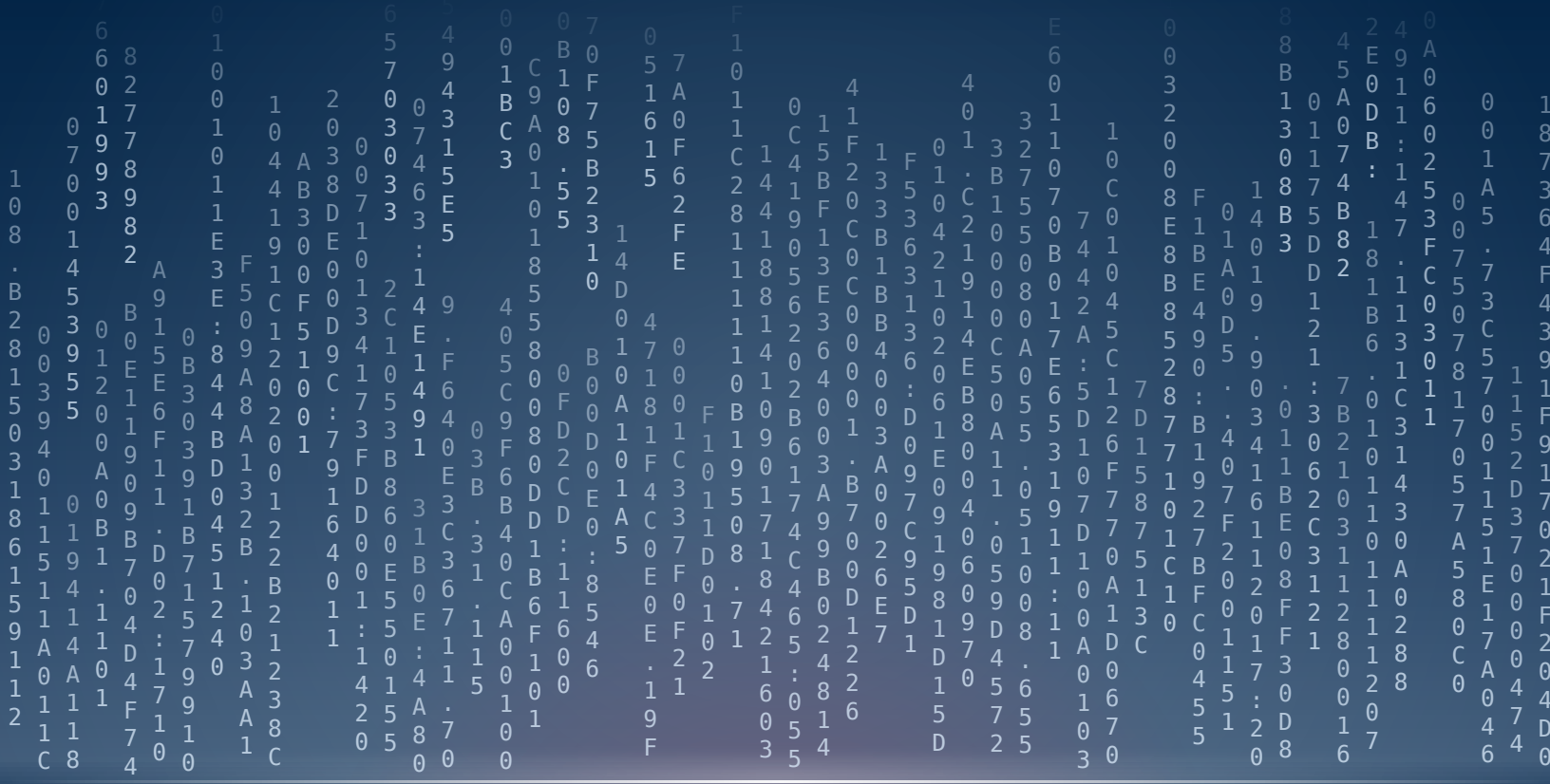
SPECIAL COMPETITIVE STUDIES PROJECT

OCTOBER 2026



China's Adversarial AI Distillation of U.S. Frontier Models

Lara Jolicoeur | Evan Armstrong

Table of Contents

China's Adversarial AI Distillation of U.S. Frontier Models 2

Why Chinese Firms Distill—and How They Do It 4

How Extracted Capabilities Reach the Chinese State 7

The Stakes: Surveillance, Influence, and Decision Advantage 14

Raising the Cost of Extraction 17

Protected Innovation Is the Advantage 19

China's Adversarial AI Distillation of U.S. Frontier Models

In 1949, the Soviet Union tested an atomic bomb one to two years sooner than American planners had expected—not because Soviet physics had caught up, but because the design had walked out of Los Alamos thanks to the spy network run by Julius and Ethel Rosenberg.[1] Soviet espionage compressed Russia's timeline for building the bomb and lowered the cost. Chinese AI firms are pursuing a similar strategy today, using a technique called "distillation" to build on AI advancements made by their American competitors and reduce the time and cost it takes them to catch up to the AI frontier. Unlike the Soviets who recruited spies to steal U.S. nuclear secrets, Chinese labs are allegedly deploying tens of thousands of fraudulent accounts to query American frontier models at an industrial scale. This 'adversarial distillation' has consequences for U.S. competitiveness and national security.

The Chinese Communist Party has prioritized the development of indigenous Chinese AI capabilities, which is seen as essential to the country's future economic prosperity and military power. For its part, the U.S. Government has enacted export controls on advanced semiconductors to protect the U.S. lead in AI and restrict Chinese firms' ability to build frontier models.[2] Distillation—training a less-capable 'student' AI model by having it learn from a more-capable 'teacher' model[3]—offers a way around U.S. export controls. Distillation is a common practice amongst AI developers; U.S. labs use it internally to improve their own models.[4]

What sets Chinese labs apart is not the practice of distillation itself, but the scale of their effort and the lengths they go to mask their activities. In 2026, OpenAI accused

---

[1] Klaus Fuchs, Atomic Heritage Foundation, National Museum of Nuclear Science & History (last accessed 2026).

[2] 国务院, 国务院关于印发新一代人工智能发展规划的通知, China Law Translate (2017); 国务院, 国务院关于深入实施"人工智能+"行动的意见, 中华人民共和国生态环境部 (2025); 87 Fed. Reg. 62,186, Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification, U.S. Department of Commerce, Bureau of Industry and Security at 62,186-62,187 (2022).

[3] Jianping Gou, et al., Knowledge Distillation: A Survey, International Journal of Computer Vision vol. 129 at 1789–1819 (2021).

[4] Model Distillation in the API, OpenAI (2024). One U.S. frontier AI developer defines 'adversarial distillation' as: "the systematic and unauthorized use of one model's outputs or reasoning to help train, reproduce, or improve another model." Disrupting a Coordinated Model-distillation Campaign, OpenAI (2026).

DeepSeek of distilling its models, and Anthropic leveled similar allegations at DeepSeek, MiniMax, Moonshot, and Alibaba, describing "industrial-scale campaigns" to extract Claude's capabilities.[5] These campaigns targeted new U.S. models immediately after they released, as shown by Anthropic's claim that MiniMax redirected nearly half its traffic within 24 hours of a new Claude release.[6] In July 2026, the Office of Science and Technology Policy said Moonshot distilled Anthropic's Fable 5 to build Kimi K3.[7] OpenAI also disclosed that it had disrupted a coordinated campaign to extract the protected reasoning behind its models, attributing a large number of the activity to individuals associated with Moonshot.[8] Chinese AI labs extensively employ so-called 'transfer stations'—websites providing proxy and application programming interface (API) access to American frontier models—to mask their distillation efforts, according to Chinese media reports.[9] Additionally, the National Security Agency, along with the Cybersecurity and Infrastructure Security Agency and Federal Bureau of Investigation, released a joint Cybersecurity Advisory in September 2026 expanding the distillation accusations to other Chinese firms StepFun and Z.AI.[10]

China's AI labs are reaping the benefits of distillation in the form of improved performance without having to make the same expensive investments in compute and training as their American competitors. AI performance benchmark gains across 2024 and early 2025 reflect both real engineering advances by Chinese labs and convergence accelerated by adversarial distillation, as suggested by Alibaba's confirmed use of GPT-4 as a black-box teacher.[11] Recent benchmarks of aggregated capabilities show

---

[5] Updated Stakes for American-Led, Democratic AI, OpenAI (2026); Detecting and Preventing Distillation Attacks, Anthropic (2026); Sarah Heck, Illicit Access to American AI Models by Alibaba-Affiliated Operators, Anthropic at 1 (2026).

[6] Detecting and Preventing Distillation Attacks, Anthropic (2026).

[7] Michael Kratsios, Statement on Moonshot AI and Kimi K3, X (2026).

[8] Disrupting a coordinated model-distillation campaign, OpenAI (2026).

[9] Cyber Security & Infrastructure Security Agency, et al., China-Based Artificial Intelligence Companies Conducting Industrial-Scale Distillation Campaigns Against U.S. AI Companies, Cyber Security & Infrastructure Security Agency (2026); Buy What It Can, Steal What It Must: China's Campaign to Acquire Frontier AI Capabilities, U.S. House of Representatives Select Committee on the Chinese Communist Party at 14 (2026); Daniel Remler & Ben Hayum, Adversarial Distillation: China's Campaign to Extract American AI Capabilities, Center for a New American Security (2026); Detecting and Preventing Distillation Attacks, Anthropic (2026).

[10] Cyber Security & Infrastructure Security Agency, et al., China-Based Artificial Intelligence Companies Conducting Industrial-Scale Distillation Campaigns Against U.S. AI Companies, Cyber Security & Infrastructure Security Agency (2026).

[11] Luke Emberson, Chinese AI Models Have Lagged the US Frontier by 7 Months on Average Since 2023, Epoch AI (2026); Daniel Remler & Ben Hayum, Adversarial Distillation: China's Campaign to Extract American AI Capabilities, Center for a New American Security (2026); Chengyu Wang, et al., EasyDistill: A

Anthropic's Claude Fable 5 scoring 163, while Kimi K3 scores 158—just five points behind Fable 5.[12] Although it is difficult to accurately estimate the compute and training costs incurred by Chinese AI labs, AI distillation is almost certainly more cost-efficient to train a student model compared to annotating additional data directly.[13]

And it is not just China's AI industry that is benefiting. Documented evidence shows DeepSeek and Qwen integrated into the People's Liberation Army (PLA) and Ministry of Public Security (MPS).[14] MiniMax and Moonshot lack comparable evidence of direct adoption, though personnel, financing, and ownership ties create credible pathways to the state.[15] The result is strengthening the PRC's military, security, and intelligence ecosystems. And because Chinese models are open-weight and available for anyone to download and fine-tune, the latest Chinese AI models are already proliferating to hostile state and non-state actors.[16] The overall impact is that the advantage that the United States seeks as it fields frontier AI is being eroded.[17]

Why Chinese Firms Distill—and How They Do It

Beijing's pursuit of global AI leadership and open-source diffusion gives Chinese firms a strong incentive to acquire frontier capabilities quickly, making adversarial distillation an attractive shortcut around the compute, cost, and research burden of building them independently.[18] The State Council's 2017 New Generation Artificial Intelligence

---

Comprehensive Toolkit for Effective Knowledge Distillation of Large Language Models, arXiv at Tbl. 1 (2025).

[12] Epoch Capabilities Index, Epoch AI (last accessed 2026).

[13] Junmo Kang, et al., Distill or Annotate? Cost-Efficient Fine-Tuning of Compact Models, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) at 1 (2023).

[14] Sunny Cheung & Kai-shing Lau, DeepSeek Use in PRC Military and Public Security Systems, Jamestown Foundation at 23-29 (2025).

[15] 管理团队, MiniMax (2026); Cheng Yu, State-Backed Investors Eye Nation's Top AI Startups, China Daily (2026); 雷晨, Kimi多了一堆"国家队"股东！20亿美元融资已到尾声, 21经济网 (2026); 中国移动发布MoMA平台：一站式调用超300款模型，单位Token成本压降30%以上, 新华网 (2026).

[16] Disrupting Malicious Uses of Our Models: An Update, February 2026, OpenAI at 1-35 (2026); Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign, Anthropic at 3, 6-14 (2025); Global AI Adoption in 2025: A Widening Digital Divide, Microsoft AI Economy Institute at 10-11 (2026); Report: Early Terrorist Adoption of Generative AI, Tech Against Terrorism at 2-8 (2023).

[17] Buy What It Can, Steal What It Must: China's Campaign to Acquire Frontier AI Capabilities, U.S. House of Representatives Select Committee on the Chinese Communist Party at 14-17 (2026); Detecting and Preventing Distillation Attacks, Anthropic (2026).

[18] To inform the analysis, we relied on public English- and Chinese-language sources, placing particular emphasis on materials produced by the Chinese government, Chinese research laboratories, and Chinese

Development Plan [新一代人工智能发展规划] set a three-step timeline: keeping pace with the world's advanced AI level by 2020; major breakthroughs in basic AI theory by 2025, with some technologies and applications reaching world-leading levels; and by 2030, world-leading AI theory, technology, and applications overall, making China the world's primary AI innovation center.[19] It also listed "open source and openness" [开源开放] as a key principle, alongside technology-led development, systematic planning, and market-driven development. The plan's successor, the "AI+" Opinion ["人工智能+"行动的意见], reaffirms these priorities.[20] But U.S. export controls imposed in 2022 bar PRC firms from legally acquiring the advanced compute needed to build frontier capabilities domestically.[21] Adversarial distillation lets Chinese firms bypass much of the compute, cost, research, safety testing, and data development required to build frontier models from scratch, while open-weight releases spread those gains across China's broader AI ecosystem.[22]

Beyond these incentives, China's legal regime supports adversarial distillation by enabling the transfer of extracted U.S. AI capabilities to the state. The Cyberspace Administration of China's (CAC) 2023 Interim Measures [生成式人工智能服务管理暂行办法], the primary regulation governing generative AI, contains only one clause on training data—requiring a "legal source" [使用具有合法来源的数据和基础模型]—never defines the term, and exempts services not yet offered to the Chinese public.[23] Chinese legal commentary reinforces the permissiveness: analyses circulating after DeepSeek's release argues trade-secret claims are weak and urges against legal restriction, while acknowledging copyright and unfair-competition risk.[24]

---

universities, as well as threat intelligence reports from U.S. technology companies. Wherever possible, we distinguish between documented distillation actions, allegations from U.S. providers, and assessed distillation inferred from Chinese lab announcements.

[19] 国务院, 国务院关于印发新一代人工智能发展规划的通知, China Law Translate (2017).

[20] 国务院, 国务院关于深入实施"人工智能+"行动的意见, 中华人民共和国生态环境部 at §III(7), §III(11) (2025).

[21] 87 Fed. Reg. 62,186, Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification, U.S. Department of Commerce, Bureau of Industry and Security at 62,186-62,187 (2022).

[22] Detecting and Preventing Distillation Attacks, Anthropic (2026); Ngor Luong, Two Loops: How China's Open AI Strategy Reinforces Its Industrial Dominance, U.S.-China Economic and Security Review Commission at 11, 15 (2026).

[23] 中国网信网, 生成式人工智能服务管理暂行办法, 中华人民共和国国家互联网信息办公室, arts. 2-7 (2023).

[24] 由DeepSeek看AI大模型研发中知识蒸馏行为的合法性, 安全内参 (2025); 傅宏宇, et al., 模型蒸馏无需受法律限制，技术创新不宜过早下结论, 安全内参 (2025).

Figure 1: Alleged Capabilities Distilled from Frontier Models

ENTITY ACCUSED	DeepSeek	Moonshot AI	MiniMax	Alibaba / Qwen
ACCUSED BY	Anthropic	Anthropic	Anthropic	Anthropic | OpenAI | U.S. Govt
COMPANY TARGETED	Anthropic Claude	Anthropic Claude	Anthropic Claude	Anthropic & OpenAI Claude & GPT
SCALE OF CLAUDE EXCHANGES	150,000+ Claude exchanges	3.4 million+ Claude exchanges	13 million+ Claude exchanges	28.8 million+ Claude exchanges via almost 25,000 fraudulent accounts (April 22 – June 5, 2026)
CAMPAIGN TARGET (FOCUS)	Reasoning capabilities and reasoning traces; generated censorship-safe alternatives to politically sensitive queries	Agentic reasoning, coding, and computer vision capabilities.	Agentic coding and tool orchestration.	High-value capabilities including agentic reasoning, software engineering, and long-horizon tasks.
KEY DETAILS	• Generated more than 150,000 Claude exchanges • Used synchronized, load-balanced traffic designed to increase throughput and avoid detection • Targeted reasoning capabilities and reasoning traces • Used Claude to generate censorship-safe alternatives to politically sensitive queries	• Generated more than 3.4 million Claude exchanges • Conducted through hundreds of fraudulent accounts • Focused on extracting agentic reasoning, coding, and computer vision capabilities	• Conducted the largest February campaign, generating more than 13 million Claude exchanges • Targeted agentic coding and tool orchestration • Redirected nearly half of its active distillation traffic to a newly released Claude model within 24 hours	• In June, operators affiliated with Alibaba and Alibaba Qwen generated more than 28.8 million Claude exchanges • Through almost 25,000 fraudulent accounts between April 22 and June 5, 2026 • Targeted high-value capabilities including agentic reasoning, software engineering, and long-horizon tasks • Allegedly supplied the Nvidia GPU cluster (~20,000 chips, reportedly incl. restricted Blackwell) used to train Moonshot's Kimi K3 — Alibaba denies

Source: Anthropic, Reuters, OpenAI[25]

With that legal and academic groundwork in place, transfer stations convert prohibited access to U.S. frontier models into scalable adversarial-distillation infrastructure by masking Chinese users, bypassing geographic and Know-Your-Customer (KYC) controls, and distributing extraction traffic across thousands of fraudulent accounts. These gray-market API relays route requests through third-party accounts, while fraudulent identities and payment cards, paid overseas account holders, and commercial proxies evade provider location, identity, and payment restrictions.[26] Anthropic described the result as a "hydra cluster": activity distributed across provider

[25] Detecting and Preventing Distillation Attacks, Anthropic (2026); Karen Freifeld, Anthropic Says Alibaba Illicitly Extracted Claude AI Model Capabilities, Reuters (2026); OpenAI Says China's DeepSeek Trained Its AI by Distilling U.S. Models, Memo Shows, Reuters (2026); Updated Stakes for American-Led, Democratic AI, OpenAI (2026).

[26] Cyber Security & Infrastructure Security Agency, et al., China-Based Artificial Intelligence Companies Conducting Industrial-Scale Distillation Campaigns Against U.S. AI Companies, Cyber Security & Infrastructure Security Agency (2026); Buy What It Can, Steal What It Must: China's Campaign to Acquire Frontier AI Capabilities, U.S. House of Representatives Select Committee on the Chinese Communist Party at 14-15 (2026); Daniel Remler & Ben Hayum, Adversarial Distillation: China's Campaign to Extract American AI Capabilities, Center for a New American Security (2026); Zeyi Yang & Matt Burgess, How People in China Keep Outsmarting Anthropic's Geolocation Restrictions, WIRED (2026).

accounts and cloud services so the campaign survives detection and account bans.[27] Across its February and June 2026 disclosures, Anthropic attributed at least 44.8 million Claude exchanges across roughly 49,000 fraudulent accounts to four Chinese AI firms, as detailed in Figure 1 above.[28] Established access lets firms systematically harvest outputs to train competing models without authorization.[29] One such site, CloseAI, advertises high-volume access and "model-distillation training," listing Alibaba, Tencent, Baidu, Tsinghua University, and Peking University as customers, although public evidence does not establish which services those entities used.[30] The campaign is industrial. The tooling is commercial. Figure 2 depicts this transfer-station pathway.

Figure 2: Depiction of Transfer Stations

Source: The National Security Agency, Cybersecurity and Infrastructure Security Agency, and Federal Bureau of Investigation[31]

How Extracted Capabilities Reach the Chinese State

China's legal architecture converts DeepSeek, MiniMax, Moonshot AI, and Alibaba's alleged adversarial distillation into a mechanism for state access. Military-Civil Fusion seeks to break down barriers between civilian research, commercial sectors, and the defense industrial base.[32] The National Defense Mobilization Law lets the state

[27] Detecting and Preventing Distillation Attacks, Anthropic (2026).
[28] Detecting and Preventing Distillation Attacks, Anthropic (2026); Karen Freifeld, Anthropic Says Alibaba Illicitly Extracted Claude AI Model Capabilities, Reuters (2026).
[29] Detecting and Preventing Distillation Attacks, Anthropic (2026); Buy What It Can, Steal What It Must: China's Campaign to Acquire Frontier AI Capabilities, U.S. House of Representatives Select Committee on the Chinese Communist Party at 14 (2026).
[30] CloseAI, CloseAI (last accessed 2026).
[31] Cyber Security & Infrastructure Security Agency, et al., China-Based Artificial Intelligence Companies Conducting Industrial-Scale Distillation Campaigns Against U.S. AI Companies, Cyber Security & Infrastructure Security Agency (2026).
[32] 国务院办公厅关于推动国防科技工业军民融合深度发展的意见, 中华人民共和国中央人民政府 (2017).

requisition civilian technology when national security is deemed threatened,[33] the National Intelligence Law obligates all organizations and citizens to support intelligence work,[34] and the Cybersecurity Law compels local data storage, access, and technical support.[35] Compounding these policies, all four firms release their models mostly as open weights, meaning the underlying technology can be freely copied, modified, and redistributed by any party—including the Chinese government.[36] Thus, gains from one model spread across firms rather than staying with a single developer.

Capability Convergence

Adversarial distillation of U.S. frontier models is likely a significant and underweighted contributor to the capability gains of Chinese models released from 2025 onward. While its public documentation cannot be established for any individual model, one clearly documented instance nonetheless exists. An Alibaba Cloud paper disclosed using OpenAI's GPT-4 as a black-box teacher to supervise its DistilQwen models, including the DistilQwen2.5 series,[37] which was released in April 2025 as lightweight open weights anyone can download and redeploy.[38] Using GPT-4's outputs this way is a textbook case of distillation, violating OpenAI's Business Terms, which since 2023 have prohibited the use of OpenAI model outputs to train competing models.[39] Thus, Alibaba Cloud's use of GPT-4 as a teacher model is a clear example of AI distillation. Alibaba Cloud's case is only one publicly documented example of a Chinese lab distilling a U.S. frontier model's outputs, and U.S. frontier labs believe the practice extends well beyond it.

China's AI frontier advanced sharply across 2024 and early 2025, and documented PRC state adoption accelerated once domestic models neared U.S. capability. The advance reflected both genuine engineering achievement—most visibly DeepSeek's architectural and training-efficiency innovations—and convergence accelerated by distillation. Our analysis leverages Epoch AI's Epoch Capabilities Index (ECI), which aggregates more

[33] 受权发布丨中华人民共和国国防动员法, 新华网 (2026).

[34] 中华人民共和国国家情报法, 全国人民代表大会 (2017).

[35] 全国人民代表大会常务委员会关于修改《中华人民共和国网络安全法》的决定, 国际科技创新中心 (2025).

[36] Ngor Luong, Two Loops: How China's Open AI Strategy Reinforces Its Industrial Dominance, U.S.-China Economic and Security Review Commission at 11 (2026).

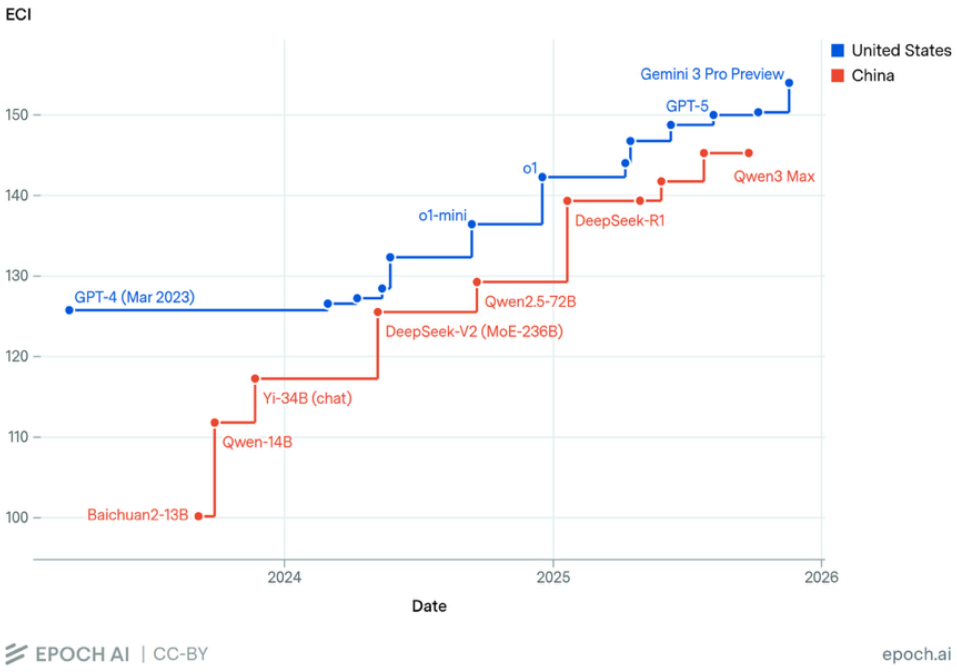
[37] Chengyu Wang, et al., EasyDistill: A Comprehensive Toolkit for Effective Knowledge Distillation of Large Language Models (2025).

[38] Alibaba-pai, DistilQwen2.5-R1 Series: Advanced Reasoning Models, Hugging Face (last accessed September 2026).

[39] Business Terms, OpenAI (2023).

than 50 benchmarks into one longitudinal measure.[40] As Figure 3 shows, China's frontier rose sharply in 2024 with DeepSeek-V2 and Qwen2.5-72B, then jumped with DeepSeek-R1 in January 2025 to approximately three points of the U.S. frontier.[41] State adoption followed that jump: Chinese large language models had been publicly available since 2023, but documented government integrations cluster in 2025 and 2026.

Figure 3: Epoch Capability Index (ECI) Comparison Between Chinese and U.S. Models



Source: Epoch AI[42]

The ECI trend admits an alternative reading—indigenous progress rather than extracted capability—and the two cannot be separated from benchmark data alone. However, the distillation allegations, Alibaba's own disclosure, and the scale of the account infrastructure described above likely make distillation a plausible contributor.

[40] Epoch AI, ECI Documentation (last accessed 2026).
[41] Luke Emberson, Chinese AI Models Have Lagged the US Frontier by 7 Months on Average Since 2023, Epoch AI (2026).
[42] Luke Emberson, Chinese AI Models Have Lagged the US Frontier by 7 Months on Average Since 2023, Epoch AI (2026).

Figure 4 below shows PLA and MPS use of and research on DeepSeek and Qwen from 2025 onward.

Documented State Adoption and Research: DeepSeek and Qwen

Figure 4: Selected Examples of Chinese Government-Linked Use of and Research on Domestic AI Models

System / Patent	Chinese Government Affiliation	AI Model	Capability Description
Integrated Intelligence, Command and Execution platform[43]	Hytera (Partially state owned and a vendor to public security bureaus across China)[44]	DeepSeek Mixture-of-Experts (MoE) architecture, run as a dual engine alongside existing generative pre-trained transformer (GPT) models	Runs three police modules: automated transcription of narrowband radio traffic into incident logs; live dispatch files linked to body-cam feeds with facial-recognition checks; and officers' spoken queries converted into database searches and pushed response plans.
National University of Defense Technology (NUDT) political education deployment[45]	People's Liberation Army	DeepSeek	Deployed locally with physical isolation from the internet for PLA political-ideological education work; the patent describes accompanying content-review mechanisms for security.
A method and system for controlling unmanned platforms using	Military Intelligence Research Institute of PLA Academy of Military Sciences	Unspecified Qwen model	A voice-and-gesture system for controlling unmanned platforms: tracked hand gestures route commands directly to the platform, and

[43] 海能达, DeepSeek遇见海能达，"AI警察"增持智能情指行，助力公安大部制改革提质增效, 东方财富网 (2025).

[44] Samuel Rubenfeld & Edmund Xu, Subsidiaries of Chinese Firms Indicted in Trade-Secret Dispute and Subject to Procurement Ban Won U.S. Government Contracts, Kharon (2022); Conor Healy, Hytera Lies to FCC About "No Ties to the Chinese Communist Party," IPVM (2021).

[45] 赵亮 and 梁友麟, DeepSeek本地化部署赋能思想政治教育初探, 国防科技大学 (2026).

multi-modal fusion of voice and gestures[46]			a wake-up gesture triggers speech-to-text command recognition.
A Method and System for Analyzing the Spread of False Information on the Internet Based on a Large Model[47]	School of Information and Communication, National University of Defense Technology Chinese PLA	Qwen2.5-0.5B-Instruct model	Trains many small models—each simulating a distinct network user—distributes them across network nodes, seeds misinformation at one node, and tracks how it spreads.

Alibaba and DeepSeek models are deployed across the PRC's defense and security apparatus. Two features matter for distillation. First, these are open-weight models run locally—NUDT's deployment is physically isolated from the internet—so adoption requires no ongoing relationship with the developer and leaves no access to revoke. Second, the capabilities in use are precisely those distillation transfers most effectively: transcription, retrieval, instruction following, multilingual analysis, and content generation. Gains extracted by Chinese firms from U.S. frontier models therefore propagate into these systems by default, through ordinary model releases.

In the research setting, the patent titled "A Method and System for Analyzing the Spread of False Information on the Internet Based on a Large Model" runs a separate Qwen2.5-0.5B-Instruct copy for each simulated network user to model how seeded misinformation propagates.[48] This could give PLA researchers a cheap way to test which messages influence audiences before deployment.[49] The risk lies not in any single model's sophistication but in running cheap models at scale—a 0.5B-parameter model

[46] CN121523554A, 一种语音、手势多模融合的无人平台操控方法及系统, Military Intelligence Research Institute, Academy of Military Sciences of the Chinese People's Liberation Army (2026).

[47] CN120567706B, 基于大模型的ai虚假信息网络传播场景分析方法及系统, School of Information and Communication, National University of Defense Technology of the Chinese People's Liberation Army (2026).

[48] CN120567706B, 基于大模型的ai虚假信息网络传播场景分析方法及系统, School of Information and Communication, National University of Defense Technology of the Chinese People's Liberation Army (2026).

[49] Nathan Beauchamp-Mustafaga, et al., Dr. Li Bicheng, or How China Learned to Stop Worrying and Love Social Media Manipulation: Insights into Chinese Use of Generative AI and Social Bots from the Career of a PLA Researcher, RAND Corporation at 41-46 (2024).

runs on a consumer laptop and can be instantiated thousands of times on domestic hardware.

On the operational side, Hytera—a communications company supplying public security bureaus across China—announced in February 2025 that it had built a DeepSeek mixture-of-experts model into its "Intelligence, Command, Action" platform as a dual engine alongside an existing GPT system.[50] The three police functions it powers rely on transcription, retrieval, data analysis, and instruction following—precisely the task-specific capabilities distillation improves.[51] DeepSeek reached this platform within weeks of its release.[52] The integration of DeepSeek adds no new collection. Instead, it removes the bottleneck between what existing systems capture and what an officer can search, match, and act on.

Beyond public security, distillation could shorten the path from U.S. frontier innovation to PLA operational use. Procurement records already show PLA demand for DeepSeek-enabled systems supporting cyberattack training, psychological warfare, military analysis, and unmanned-vehicle command and control.[53] Defense-university survey literature reviews adapting large language models for command-and-control, situational awareness, intelligence analysis, and cognitive-domain operations, and Chinese work now proposes formal frameworks to evaluate military large models.[54]

Base 311, also known as Unit 61716, shows how those capabilities could reach an operational mission: affiliated researchers have explored AI systems that analyze public opinion, generate posts, coordinate accounts, and adapt influence campaigns as

---

[50]海能达, DeepSeek遇见海能达，"AI警察"增持智能情指行，助力公安大部制改革提质增效, 东方财富网 (2025); 公共安全解决方案 [Public Safety Solutions], Hytera (archived July 24, 2026); How ChatGPT and Our Foundation Models Are Developed, OpenAI (last accessed 2026); Alec Radford, et al., Improving Language Understanding by Generative Pre-Training, OpenAI at 1, 3 (2018).

[51] Detecting and Preventing Distillation Attacks, Anthropic (2026); Model Distillation in the API, OpenAI (2024).

[52] 海能达, DeepSeek遇见海能达，"AI警察"增持智能情指行，助力公安大部制改革提质增效, 东方财富网 (2025); DeepSeek-R1 Release, DeepSeek (2025).

[53] Cole McFaul & Sam Bresnick, The National Security Case for Limiting China's Access to Advanced U.S. Compute: Evidence from PLA Procurement Documents, Emerging Technology Observatory (2026); Emelia Probasco, et al., China's Military AI Wish List: Command, Control, Communications, Computers, Cyber, Intelligence, Surveillance, Reconnaissance, and Targeting (C5ISRT), Center for Security and Emerging Technology at 31-32 (2026).

[54] 高志强, et al., 大模型技术的军事应用综述, 南京航空航天大学学报 (2024); 蔡磊, et al., 大模型在军事垂直领域的应用, 船舶科学技术; 军队军工, AIGC军事大模型评估体系框架研究, 安全内参 (2025); 军队军工, 面向网络空间认知战的大语言模型：技术与挑战, 安全内参 (2025).

discourse changes.[55] Leaked GoLaxy documents list Unit 61716 as a customer, indicating an existing commercial pathway into the unit.[56] If distillation improves the reasoning, multilingual analysis, and content-generation capabilities of Chinese models, those gains could flow into Unit 61716's political warfare mission against Taiwan—and across other PLA missions whose recipient units are not publicly identifiable—without the unit conducting any distillation itself.

The open literature shows the same ambition: a 2025 survey of large language models for cyberspace cognitive warfare describes mass-producing targeted false intelligence to degrade an adversary's decision-making,[57] and one vendor already markets a domestic "military large model" bundling an automatic model-distillation plug-in to spin up mission-specific models cheaply.[58]

Indirect Pathways: MiniMax and Moonshot

Unlike DeepSeek and Qwen, MiniMax has no documented PRC government adoption, but its leadership pipeline from SenseTime links it indirectly to China's surveillance ecosystem. Three senior leaders, including its chief executive, previously worked at SenseTime,[59] which the U.S. Government sanctioned for its facial-recognition technology designed to identify Uyghur ethnicity and its involvement with the Chinese military.[60] MiniMax's consumer services collect the relevant data: Xingye and Talkie are AI-companion platforms,[61] and Hailuo AI processes facial geometry and movement to

[55] Ryan Fedasiuk, et al., Harnessed Lightning: How the Chinese Military Is Adopting Artificial Intelligence, Center for Security and Emerging Technology at 53, 63-64 (2021); Nathan Beauchamp-Mustafaga, et al., Dr. Li Bicheng, or How China Learned to Stop Worrying and Love Social Media Manipulation: Insights into Chinese Use of Generative AI and Social Bots from the Career of a PLA Researcher, RAND Corporation at 40-46 (2024).

[56] Brett Goldstein & Brett Benson, The GoLaxy Documents: Inside One Chinese Company's AI-Driven Influence Machine, Vanderbilt University Institute of National Security at 379 (2025).

[57] 军队军工, 面向网络空间认知战的大语言模型：技术与挑战, 安全内参 (2025).

[58] 渊亭科技, 八大能力体系，70+应用场景！渊亭科技推出国内首个军事大模型, 搜狐 (2023).

[59] 管理团队, MiniMax (2026).

[60] Bureau of Industry and Security, U.S. Department of Commerce, Addition of Certain Entities to the Entity List, 84 Fed. Reg. 54,002 (2019); U.S. Department of Defense, Entities Identified as Chinese Military Companies Operating in the United States in Accordance with Section 1260H of the William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021 (2021); Office of Foreign Assets Control, Treasury Sanctions Perpetrators of Serious Human Rights Abuse on International Human Rights Day, U.S. Department of the Treasury (2021).

[61] About, Talkie (2024), 星野,XingyeAI (2026).

generate realistic human video.[62] Its privacy policy permits disclosure to government requests,[63] and China's intelligence and cybersecurity laws could compel transfer to the state.

Similarly, Moonshot's state-backed financing and distribution through China Mobile's Model Management (MoMA) platform create indirect pathways, despite there being no evidence of direct PLA or MPS adoption. In 2026, Moonshot added the Beijing AI Industry Investment Fund, Guozhitou—manager of China's roughly RMB 60 billion national AI fund—and state-owned China Mobile to its shareholder roster.[64] China Mobile has incorporated Kimi into MoMA, a platform expressly designed for government-affairs, financial, industrial, health, and education applications.[65] MoMA, therefore, gives capabilities that Anthropic alleges Moonshot distilled from Claude, a state-owned distribution channel into government-oriented systems.[66]

The Stakes: Surveillance, Influence, and Decision Advantage

Taken together, these pathways could turn U.S. frontier-model capabilities into both deployable PRC state capacity and a force multiplier for acquiring foreign technology. DeepSeek's appearance in PLA procurements and public-security systems shows how U.S.-developed reasoning, coding, and analytical capabilities could migrate into surveillance, influence, and transnational-repression workflows.[67] These same capabilities could accelerate China's absorption of foreign technological capabilities by helping actors identify targets, interpret technical materials, and reverse engineer acquired intellectual property,[68] positioning U.S. frontier innovation as both an

---

[62] MiniMax Hailuo 2.3: A New Level of Complex Video Performance & Media Agent, MiniMax (2025); Hailuo AI Privacy Policy, MiniMax (2025).
[63] MiniMax App and Web Privacy Policy, MiniMax (2025); Hailuo AI Video Privacy Policy, Hailuo AI (2025).
[64] Cheng Yu, State-Backed Investors Eye Nation's Top AI Startups, China Daily (2026); 雷晨, Kimi多了一堆"国家队"股东！20亿美元融资已到尾声, 21 经济网 (2026).
[65] 中国移动发布MoMA平台：一站式调用超300款模型，单位Token成本压降30%以上, 新华网 (2026).
[66] Detecting and Preventing Distillation Attacks, Anthropic (2026); Chinese AI Companies 'Distilled' Claude to Improve Own Models, Anthropic Says, Reuters (2026).
[67] Sunny Cheung & Kai-shing Lau, DeepSeek Use in PRC Military and Public Security Systems, China Brief, The Jamestown Foundation at 23-29 (2025); Detecting and Preventing Distillation Attacks, Anthropic (2026).
[68] Detecting and Preventing Distillation Attacks, Anthropic (2026); Findings of the Investigation into China's Acts, Policies, and Practices Related to Technology Transfer, Intellectual Property, and Innovation Under Section 301 of the Trade Act of 1974, Office of the U.S. Trade Representative at 14-18 (2018).

appropriated end product and a tool for acquiring more advancements.[69] SCSP's Offset-X Evolved strategy rests future U.S. warfighting on a triad of advantages in sensing, AI, and autonomy. Adversarial distillation strikes the AI leg of that triad directly, converting an American advantage into a shared one.[70]

Beyond state capacity, adversarial distillation threatens the ability of providers and governments to control how advanced capabilities are used once they are embedded in open-weight Chinese models.[71] Anthropic warns that illicitly distilled models may shed safeguards, letting capabilities once constrained by U.S. providers move into military, intelligence, surveillance, and influence workflows with fewer protections.[72] Once released as open weights, those capabilities become durable and transferable, beyond provider visibility or revocation. DeepSeek has gained traction in Russia and Iran, and terrorists and violent extremists have begun experimenting with generative AI, illustrating that Chinese open models widen access to capabilities first developed at the U.S. frontier.[73]

As Figure 5 shows, PRC-linked actors already apply provider-hosted frontier models across surveillance, influence, and cyber workflows. Distillation would convert those monitored, revocable gains into durable Chinese-controlled capacity no U.S. provider can observe, constrain, or revoke. OpenAI's "Peer Review" case illustrates how frontier models compress the analytical labor of mapping networks and producing tailored content; transferred into a PRC-controlled model, that capability could scale surveillance and coercion against U.S.-based diaspora and policy communities without the safeguards that currently limit it.

---

[69] Findings of the Investigation into China's Acts, Policies, and Practices Related to Technology Transfer, Intellectual Property, and Innovation Under Section 301 of the Trade Act of 1974, Office of the U.S. Trade Representative at 14-18, 62-147, 153-173 (2018); Detecting and Preventing Distillation Attacks, Anthropic (2026).

[70] Offset-X Evolved: A Technology-Powered Strategy for the U.S. Military, Special Competitive Studies Project (2025).

[71] Disrupting Malicious Uses of Our Models: An Update, February 2026, OpenAI at 26-27 (2026); Detecting and Preventing Distillation Attacks, Anthropic (2026).

[72] Detecting and Preventing Distillation Attacks, Anthropic (2026).

[73] Global AI Adoption in 2025: A Widening Digital Divide, Microsoft AI Economy Institute at 10-11 (2026); Maggie Engler, Considerations of the Impacts of Generative AI on Online Terrorism and Extremism, Global Internet Forum to Counter Terrorism at 8-9 (2023); Report: Early Terrorist Adoption of Generative AI, Tech Against Terrorism (2023).

Figure 5: Findings from Anthropic, Google, and OpenAI's Threat Intelligence Reports

Case	Reported activity	Scale, significance, and limitations
OpenAI "Peer Review" (2025)[74]	Likely China-origin actors researched U.S. think tanks and officials, analyzed Uyghur protest materials, drafted dissident commentary, and developed the claimed Qianyue [千阅境外舆情AI助手] social-media monitoring tool.	Illustrates potential surveillance of demonstrations, diaspora communities, and policy networks. OpenAI attributed the activity only to likely China-origin actors.
OpenAI "Cyber Special Operations" (2026)[75]	An account linked to Chinese law enforcement described overseas intimidation, impersonated U.S. immigration officials, forged court documents, and used DeepSeek-R1, Qwen2.5, and YOLOv8 workflows for monitoring, profiling, translation, and reporting.	Reports described hundreds of personnel and thousands of fake accounts; one Japan-focused operation proceeded after ChatGPT refused assistance. OpenAI could not verify the full scope.
OpenAI and Google cyber reporting (2025)[76]	APT5- and APT15-associated accounts used AI for reconnaissance, scripting, and infrastructure profiling; Google observed similar activity by more than 20 PRC-backed groups.	Neither provider found fundamentally novel capabilities, but automation could increase operational speed and scale.
Anthropic cyber reporting (2025)[77]	A suspected Chinese actor used Claude across 12 of 14 MITRE ATT&CK tactics in a nine-month campaign against Vietnamese critical infrastructure; GTG-1002 later used Claude Code against roughly 30 entities, with Claude performing 80 to 90% of tactical work.	Shows progression from task assistance to campaign orchestration. Humans retained escalation and exfiltration decisions.

[74] Disrupting Malicious Uses of Our Models: An Update, February 2025, OpenAI at 1-35 (2025).
[75] Disrupting Malicious Uses of Our Models: An Update, February 2026, OpenAI at 1-35 (2026).
[76] Disrupting Malicious Uses of AI: June 2025, OpenAI at 1-35 (2025); Google Threat Intelligence Group, Adversarial Misuse of Generative AI, Google Cloud (2025).
[77] Threat Intelligence Report: August 2025, Anthropic at 18 (2025); Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign, Anthropic at 3, 6-14 (2025).

Finally, as the United States embeds frontier AI into national-security missions, adversarial distillation could erode its decision advantage without requiring the theft of model weights. In 2025, the Defense Department awarded AI contracts and expanded access through GenAI.mil and classified-system agreements.[78] China could erode this advantage by combining distilled reasoning, planning, translation, and tool use with Chinese data, doctrine, and mission systems—a risk that would heighten if China gained unauthorized access to restricted U.S. mission models, whose specialized workflows distillation could reveal.[79] Decision advantage—the prize SCSP has placed at the center of the AI competition for both intelligence and defense—is precisely what adversarial distillation erodes.[80]

Raising the Cost of Extraction

The recommendations below are sequenced to match the problem: give prosecutors a statute that reaches the conduct, give defenders a shared threat picture, close the access loopholes that make extraction cheap, and compete for the open tier that safeguards alone cannot protect.

Clarify the Economic Espionage Act.

Current law does not clearly establish whether systematically harvesting proprietary model outputs or reasoning traces to reproduce a model's capabilities constitutes trade-secret theft under 18 U.S.C. §§ 1831 and 1839 when done through deception or circumvention of access controls for the benefit of a foreign government.[81] The Department of Justice could be directed to issue guidance distinguishing adversarial distillation from authorized research and ordinary model use, giving prosecutors a clear

[78] U.S. Defense Department Awards Contracts to Google, Musk's xAI, Reuters (2025); The War Department Unleashes AI on New GenAI.mil Platform, U.S. Department of War (2025); Ben Finley & Matt O'Brien, U.S. Military Reaches Deals with 7 Tech Companies to Use Their AI on Classified Systems, Associated Press (2026).
[79] Anthropic, Detecting and Preventing Distillation Attacks, Anthropic (2026); U.S. House Select Committee on the Strategic Competition Between the United States and the Chinese Communist Party, Buy What It Can, Steal What It Must: China's Campaign to Acquire Frontier AI Capabilities, U.S. House of Representatives at 14-17 (2026).
[80] Intelligence Innovation: Repositioning for Future Technology Competition, Special Competitive Studies Project (2024); Memo to the President and Congress on the Adoption of Generative Artificial Intelligence for Military Purposes, in Generative AI: The Future of Innovation Power, Special Competitive Studies Project (2023).
[81] 18 U.S.C. § 1831, Economic Espionage; 18 U.S.C. § 1839, Definitions.

basis to pursue state-linked campaigns without criminalizing routine interaction with AI services.

Make government–industry threat-intelligence sharing on distillation attacks durable.

National Science and Technology Memorandum 4 directs the federal government to share information on foreign distillation campaigns with U.S. AI companies and develop defensive best practices with industry.[82] Because a presidential memorandum can be rescinded, these functions are vulnerable to changes in administration. They should remain across Administrations. OpenAI, Anthropic, and Google DeepMind already share detection information through the Frontier Model Forum.[83] The U.S. government should also support the Institute for AI Policy and Strategy's proposal for a National Institute of Standards and Technology-led AI Distillation Defense Framework setting minimum standards for access controls, detection, and response—extended beyond the frontier laboratories to smaller providers, cloud platforms, and API aggregators.[84]

Support a competitive U.S. open-weight ecosystem.

The U.S. government should support American open-weight models through federal research funding, access to compute for U.S. developers and academic institutions, and procurement preferences. Chinese open-weight releases, including the Qwen and DeepSeek models documented above, have become the default low-cost option for developers worldwide, and each adoption deepens reliance on models whose developers are subject to PRC legal-access mechanisms. Anti-distillation safeguards address extraction; they do not address the demand Chinese open models meet. A credible U.S. alternative would reduce that dependence and limit the strategic return Beijing earns from distillation-derived capabilities.

This support is consistent with the report's warning that open-weight releases diffuse capabilities irrevocably—provided the two tracks are kept distinct. American open models would sit deliberately below the protected frontier, and no open release—American or Chinese—can be kept from the PLA once published. Policymakers should

[82] Michael J. Kratsios, Memorandum on Adversarial Distillation of American AI Models, White House Office of Science and Technology Policy (2026).
[83] Shirin Ghaffary & Maggie Eastland, OpenAI, Anthropic, Google Unite to Combat Model Copying in China, Bloomberg (2026).
[84] Theo Bearman, AI Distillation Attacks: The Case for Targeted Government Intervention, Institute for AI Policy and Strategy (2026).

therefore pair open-weight support with capability evaluations before release and monitoring of U.S.-origin weights for adversarial adaptation. The objective is not to deny China open models, which is impossible, but to ensure the open tier the world builds on is American and transparent while the capabilities that confer decisive military advantage remain behind the frontier labs' hardened access controls. America's innovation power is the foundation of its competitive advantage; ceding the global open tier would cede part of that foundation.[85]

Protected Innovation Is the Advantage

The United States leads by creating frontier models; adversarial distillation enables Chinese firms to reproduce those capabilities faster and more cheaply despite the compute gap. As firms accused of distillation supply models to PLA procurement, defense-linked research, and public-security systems, U.S.-developed capabilities could diffuse across PRC military and intelligence missions beyond any provider's control—frontier models indirectly serving the PRC state without their creators' consent. In a conflict, the PLA could turn capabilities derived from America's own innovation against the United States or its allies.

To combat adversarial distillation, the United States should pair stronger safeguards with support for a competitive domestic open-weight ecosystem: access controls, threat sharing, and technical defenses to raise the cost of extraction; capable U.S. open models to reduce reliance on Chinese alternatives. Together, these measures would protect American innovation and prevent frontier models from becoming a subsidized development pipeline for the PRC.

The window for acting is the decisive decade SCSP has warned of since its founding: choices made and actions taken before 2030 will determine whether America's AI lead endures.[86] In 1949, Washington learned that a monopoly it expected to hold for a decade had already been quietly copied. Innovation alone is not an advantage. Protected innovation is. The United States should not relearn that lesson at machine speed.

[85] Generative AI: The Future of Innovation Power, Special Competitive Studies Project (2023).
[86] Mid-Decade Challenges to National Competitiveness, Special Competitive Studies Project (2022).