COMMENTARIES

Beyond the Frontier: Industrial Distillation, Inference Laundering, and China’s Asymmetric AI Proliferation

Author: Abhay A. • **Published:** October 26, 2026 • **Source:** Centre for New Age Warfare Studies



EXECUTIVE BRIEF & KEY FINDINGS

The rapid advancement of Chinese artificial intelligence capabilities raises questions about the methods through which technological advantages are acquired and replicated. Drawing on Anthropic’s September 2026 threat intelligence disclosures, this commentary examines the alleged industrial-scale distillation campaigns involving Chinese AI laboratories, focusing on automated query operations, fraudulent accounts and proxy networks used to extract capabilities from frontier models.It explores how the transfer of advanced reasoning and coding capabilities without equivalent safety safeguards could create new risks across cybersecurity, sensitive research and national security. The analysis argues that inference laundering represents an emerging dimension of asymmetric technological competition, potentially allowing actors to accelerate AI development while bypassing the costs and constraints of independent innovation. It concludes that effective responses must extend beyond semiconductor export controls to include stronger model-access governance, API monitoring and safeguards against unauthorised capability extraction.

For years, the geopolitical discourse surrounding artificial intelligence and the People's Republic of China (PRC) has relied on a neat binary: Western labs lead in foundational frontier research backed by massive capital, while Chinese labs achieve breakthroughs through indigenous scaling, algorithmic efficiency, and open-source adaptation. Recent intelligence disclosures, most notably Anthropic’s September 2026 threat intelligence report challenged this paradigm.

They describe an industrial-scale, covert ecosystem of capability extraction. Moving far beyond traditional intellectual property disputes, major PRC-based AI labs have engaged in large-scale distillation campaigns involving the systematic acquisition of frontier-model capabilities.

The Mechanics of Inference Laundering: From Open-Market Theft to Proxy Networks

In traditional strategic diffusion, states typically acquire capabilities by reverse-engineering physical hardware or cracking source code. In the realm of advanced AI, the strategic asset of highest value is frontier reasoning, specifically multi-step logical deduction, coding proficiency, and agentic tool-use. Anthropic's September 2026 disclosures describe campaigns attributed to seven China-based AI labs, including Alibaba, Moonshot AI, DeepSeek, Zhipu, Xiaomi, SenseTime, and MiniMax.
These operations went beyond conventional data scraping, employing automated query campaigns, fraudulent accounts, intermediary infrastructure, and techniques to extract reasoning traces for model training.

Alibaba's Chain-of-Thought (CoT) distillation pipeline reportedly peaked at nearly 3 million exchanges per day, launched through more than 3,500 fraudulent accounts. Anthropic observed more than 151 million exchanges attributable to Alibaba between May and July 2026.
The report also describes Moonshot and DeepSeek rerouting customer requests to Claude for capability extraction and model training.

This infrastructure functioned as a proxy-mediated capability extraction network, enabling unauthorised access to advanced models while potentially reducing the substantial time, computational resources, and expenditure required for independent model development. The significance is not simply that proprietary outputs are being copied, but that sophisticated intermediary networks can be used to acquire capabilities developed by frontier AI laboratories at scale.

The Distillation Hazard: Why Extracting Reasoning Can Weaken the Safety Wrapper

To understand why these industrial-scale distillation pipelines represent an acute defence concern, one must examine the asymmetry between model capability and alignment. When a smaller student model is trained on the outputs or reasoning traces of a heavily safeguarded teacher model, useful capabilities may transfer without the teacher's full safety training, refusal behaviour, and deployment controls being reproduced.

Anthropic reports that its own research found that the robust safeguards preventing Claude from being misused did not transfer when its models were distilled by unauthorised laboratories. This is a significant finding, although it should not be interpreted as meaning that every distilled model automatically loses all safety measures. The more precise concern is that capability transfer does not guarantee equivalent safety transfer.

Frontier labs employ multiple safeguards to reduce the risk of models generating actionable assistance for malicious exploitation. Illicit distillation can create an asymmetry in which useful reasoning, coding, and agentic capabilities are acquired without equivalent protective mechanisms being reproduced in the resulting model. Consequently, a distilled model may retain advanced technical capabilities while exhibiting weaker safeguards, potentially lowering barriers to misuse in sensitive domains.
The extent of this divergence ultimately depends on the training process and the safeguards implemented in the receiving model.

Anthropic's research also indicates that distillation can improve performance in dangerous domains, including biological and cyber-related tasks, even when the harvested exchanges contain little material specifically concerning those domains. This raises the possibility that capability extraction can produce benefits extending beyond the immediate

tasks targeted during the distillation campaign.

Strategic Divergence: PRC Cyber Targeting and Guardrail Shopping

When examining how actors navigate high-consequence domains, reported threat activity can provide important indications of deliberate capability-seeking behaviour. In the cyber domain, the evidence described in the Zhipu (Z.ai) case study points beyond passive coding assistance towards the targeted extraction and evaluation of Western frontier models and cyber capabilities. Zhipu researchers used public vulnerability datasets to develop capture-the-flag (CTF) challenges and launched distillation campaigns targeting leading US frontier models.

Zhipu initially attempted to target Anthropic's Fable model, which had strengthened cybersecurity safeguards. According to Anthropic's report, those safeguards degraded the effectiveness of Zhipu's attempts, leading its researchers to abandon Fable and switch to Opus 4.6 and another leading US model because they assessed the alternative safeguards to be weaker. Anthropic also reported that Zhipu used Opus 4.6 primarily to evaluate and grade responses produced by the other frontier model.

This behaviour indicates deliberate model selection based on perceived safeguard permissiveness. It supports the assessment that the operators sought to circumvent restrictions that impeded their capability-extraction objectives. The reported sequence of targeting, encountering safeguards, and switching models provides stronger evidence of intentional safeguard-seeking than of passive capability exploration.

Anthropic's biological-misuse case studies describe related patterns in which actors sought alternative model access when safeguards or regional restrictions impeded their objectives. Although these cases should not be treated as evidence of a common operator or identical intent, they reinforce the broader concern that safeguards can encourage attempts to find alternative access pathways rather than eliminate demand for high-consequence capabilities.

Conclusion: Defining the Strategic Intent

When viewed through a strategic framework, these recurring patterns suggest deliberate efforts to acquire advanced AI capabilities at scale. The volume of model exchanges, the use of fraudulent accounts and proxy infrastructure, the extraction of reasoning traces, and repeated attempts to bypass technical safeguards point to organised efforts to obtain capabilities developed by frontier-model providers.

The evidence is particularly significant because similar methods appear across several organisations. Alibaba used large-scale reasoning-trace distillation; Moonshot and DeepSeek reportedly routed requests to Claude and developed pipelines to extract its reasoning; Zhipu evaluated cyber capabilities while moving away from safeguards it considered stronger; and Xiaomi replayed conversations from its own models through Claude to generate training data.

Taken together, these activities suggest a broader pattern of seeking to reduce the time and cost of advanced AI development, accelerate domestic model development, and exploit weaknesses in existing access controls and safeguards. Although the report attributes the campaigns to specific PRC-based organisations with high confidence, this does not necessarily establish that every campaign was centrally directed by the Chinese state, as organisational involvement and central state direction are distinct questions.

The broader concern is that large-scale capability extraction could erode the technological advantages of frontier-model developers. It could also allow advanced capabilities to spread beyond the safety frameworks in which they were originally developed, creating risks for cybersecurity, sensitive research, data protection, and potentially national defence.

If Western policy responses focus only on traditional export controls on semiconductor hardware, they risk overlooking another important dimension of AI competition. Policymakers also need to address how access to frontier models is controlled through APIs, how abusive proxy networks are detected, how sensitive user data is protected, and how unauthorised extraction of model capabilities is prevented. AI competition increasingly depends not only on control over computing hardware but also on the ability to secure access to advanced models, detect large-scale attempts to extract their capabilities, and manage how those capabilities spread.

The views expressed are those of the authors and do not represent the views of CNAWS.

?ImageSource: ChatGPT