LLM Model Extraction at Cloud Scale: The 16 Million Query IP Theft Vector

Unofficial AI-assisted Research

Cloud Security Alliance AI Safety Initiative

2026-03-08

Key Takeaways

On February 23, 2026, Anthropic publicly disclosed that three Chinese AI companies—DeepSeek, Moonshot AI, and MiniMax—collectively issued more than 16.55 million queries to its Claude API through approximately 24,000 fraudulent accounts, with the explicit objective of extracting Claude's differentiated capabilities into competing models through a process known as knowledge distillation [1]. To the authors' knowledge, the disclosure represents the first time a major AI provider has publicly attributed large-scale model extraction activity to specific named companies, and it arrives at a moment of active legislative debate over AI chip export controls. The scale, sophistication, and geopolitical context of this incident transforms what was primarily an academic research concern into an operationally documented, policy-relevant attack class with identified actors.

Model extraction at this scale is not primarily a credential theft or access control problem in the traditional sense. Anthropic's investigation found no evidence of unauthorized access to internal systems. The attackers purchased legitimate API access—albeit through fraudulent account clusters that evaded geographic restrictions—and used that access precisely as designed, at volume, with structured prompts calibrated to transfer specific capabilities. This is an attack that exploits the fundamental design of commercial AI APIs: that responding helpfully to queries is the product. Security practitioners must therefore approach model extraction as a behavioral detection problem rather than a perimeter defense problem, and organizations that develop, deploy, or depend on proprietary AI capabilities need to reassess their threat models accordingly.

Background

What Model Extraction Is and Why It Matters

Knowledge distillation, the legitimate machine learning technique that underpins model extraction attacks, was originally developed to transfer learned capabilities from a large, computationally expensive model (the "teacher") into a smaller, more efficient model (the "student") for deployment purposes. The technique works because a large model's probability distributions over output tokens carry far more information than the raw labels in a training dataset—they encode the model's uncertainty,

generalization, and reasoning patterns in a form that can be efficiently transferred. Applied maliciously against a commercial API, the same principle allows an attacker to treat any API-exposed model as an unwitting teacher, collecting (prompt, response) pairs at volume to generate labeled training data that represents the model's behavior across targeted capability domains.

The commercial significance of this threat stems from the economics of frontier AI development. Training a frontier large language model requires billions of dollars in compute [3], years of research labor, and accumulated proprietary data, safety work, and fine-tuning that are not reflected in the base model weights but are embedded in the resulting model's behavior. An attacker who successfully extracts targeted capabilities through distillation incurs only the cost of API queries plus fine-tuning a base model—costs that academic research has demonstrated can be surprisingly low. Mindgard's January 2025 research documented task-specific extraction of GPT-3.5-Turbo at 73 percent Exact Match and 87 percent F1 similarity for approximately $50 in API costs [2]. The CSIS estimated that DeepSeek's broader training campaign—of which distillation appears to have been a component—cost approximately $6 million in total, compared to the billions invested by frontier developers [3].

The legal landscape for addressing model extraction is notably unsettled. U.S. copyright law, as clarified by the Copyright Office in January 2025, requires human authorship for protection, leaving AI-generated outputs with limited copyright standing [4]. Trade secret claims are structurally available—distillation attacks could constitute misappropriation of trade secrets embedded in model behavior—but under U.S. trade secret doctrine, enforcement may require establishing that access was unauthorized, a complex standard when API access was legitimately purchased through proxy intermediaries [14]. "Difficult" is also evaluative without specifying the relevant legal threshold or jurisdiction. The most immediately actionable remedies are contractual: service terms universally prohibit using API outputs to train competing models, and violations provide grounds for account termination and civil litigation, though cross-border enforcement against actors in non-treaty jurisdictions remains uncertain.

The Three Threat Actors: Scope and Methodology

Anthropic's disclosure documented materially different operational profiles across the three attributed actors, reflecting different stages of capability development and different target capabilities [1]. DeepSeek's campaign was the smallest in query volume at over 150,000 exchanges, focused on eliciting reasoning and rubric-based reward modeling capabilities, suggesting a targeted attempt to replicate specific components of Claude's Constitutional AI training process. Moonshot AI conducted over 3.4 million exchanges spanning agentic reasoning, tool use, coding, and computer vision, representing a broader capability acquisition effort. MiniMax's campaign was the largest, at approximately 13 million exchanges, targeting agentic coding and tool orchestration, and was uniquely notable because Anthropic detected it while still active—before MiniMax released the model being trained. When Anthropic released

a new version of Claude during the active MiniMax campaign, MiniMax pivoted within 24 hours to redirect nearly half its traffic to the new model, demonstrating operational agility consistent with institutional rather than ad hoc execution—though this timing characterization is Anthropic's own assessment and has not been independently verified.

All three actors bypassed geographic access restrictions—Anthropic prohibits commercial API access from China for legal, regulatory, and security reasons—through commercial proxy services operating "hydra cluster" architectures: networks of 20,000 or more simultaneously active fraudulent accounts that blended distillation traffic with legitimate-appearing requests. The use of commercial proxy resellers indicates that the infrastructure for large-scale API abuse is itself commoditized, reducing the technical barrier for future actors to mount comparable campaigns. Google's Threat Intelligence Group (GTIG) corroborated the general threat landscape in a simultaneous February 2026 report, disclosing that it had detected and disrupted model extraction attempts against Gemini throughout 2025, including a campaign involving over 100,000 prompts targeting reasoning replication across non-English languages [5].

Security Analysis

The Detection Gap: Why Standard Controls Failed

The most operationally significant finding from Anthropic's investigation is what was absent from the attack signature. The hydra cluster architecture produced no elevated authentication failure rates, no per-account rate limit violations, and no individually anomalous request velocities—all signals that conventional API abuse detection systems are calibrated to identify. Each fraudulent account within the cluster behaved within normal usage parameters. The anomalous pattern was only visible in aggregate, across account clusters, over time: massive volume concentrated in narrow capability domains, highly repetitive prompt structures, and content that mapped directly to what is most valuable for model training. Treblle's independent technical analysis of the incident highlighted the gap between authentication and observability—knowing who is calling tells you nothing about what they are actually doing—characterizing the incident as a failure of authenticated access controls to surface adversarial intent [6].

This creates a detection problem that is architectural rather than parametric. Adding more rate limits or stricter per-account thresholds would have minimal effect against a well-resourced actor willing to operate 20,000 accounts simultaneously. Effective detection requires behavioral fingerprinting and cross-account correlation at a level of sophistication that few organizations maintain by default for API

traffic, and which requires sustained investment in machine learning-based monitoring systems that are themselves non-trivial to build and maintain. Academic research on distillation attack evasion further complicates the picture: the LoRD (Locality Reinforced Distillation) algorithm, published in September 2024, demonstrated that policy-gradient-style extraction methods can be designed to specifically minimize query volume while maximizing capability transfer, reducing the footprint that volume-based detectors would observe [7].

The extracted model's safety alignment is a secondary but material concern. Models trained through distillation inherit their teacher's behavioral capabilities but not necessarily the underlying safety training that shaped those capabilities. Anthropic's disclosure noted this explicitly, observing that distillation campaigns that selectively target reasoning and agentic capabilities while avoiding safety-relevant training signals can produce capable models with reduced safety alignment relative to the source. This means that successful model extraction is not simply an economic harm to the original developer—it can produce less-aligned AI systems at scale, with downstream implications for the security posture of applications built on those distilled models.

Related Attack Vector: LLMjacking

Model extraction as described above is distinct from a related but separate threat class known as LLMjacking, first documented by the Sysdig Threat Research Team in May 2024 [8]. In LLMjacking attacks, adversaries steal cloud service credentials—in the documented case, exploiting CVE-2021-3129, a Laravel framework vulnerability—and use those credentials to access cloud-hosted LLM APIs across ten or more providers simultaneously, including Anthropic, OpenAI, AWS Bedrock, Azure, and Google Cloud Vertex AI. The financial objective is not capability extraction but access resale: stolen credentials are used to provision LLM access that is then sold to other parties, while the credential victim bears the compute charges. Sysdig estimated potential victim costs exceeding $46,000 per day in LLM consumption charges [8].

While LLMjacking and distillation attacks are separate in objective and technique, they share infrastructure characteristics that are relevant to enterprise detection programs. Both require large-scale, automated API interaction through proxied or stolen credentials; both are calibrated to avoid per-session anomaly signals; and both represent the exploitation of AI API surfaces for economic value. Organizations that detect credential abuse patterns consistent with LLMjacking should assess whether the same infrastructure might be used for capability extraction purposes, and vice versa.

IP Theft as Geopolitical Instrument

Anthropic's February 2026 disclosure was explicitly framed in geopolitical terms, with material implications for policy debates over AI chip export controls. The disclosure was published during active U.S. congressional deliberation over AI chip export controls, and Anthropic explicitly argued that distillation attacks "reinforce the rationale for export controls"—the premise being that chip access restrictions designed to slow frontier AI development outside U.S. jurisdiction are partially undermined if capabilities can be transferred via API distillation without chip access. The CSIS analysis characterized this dynamic as an existential competition in which the winner "will write the rules of the emerging international order" [3]. It is worth noting that Anthropic, as a party with regulatory and commercial interests in how this incident is framed, is not a neutral observer in this characterization.

Security practitioners should understand this framing not because geopolitical policy is within the scope of enterprise security programs, but because it informs the resource level and institutional backing that adversaries in this space may have available. The scale and coordination of these campaigns—24,000 fraudulent accounts, commercial proxy infrastructure, and rapid operational pivoting when the target model version changes—is consistent with institutional rather than ad hoc execution, and may indicate state-adjacent backing, though Anthropic's disclosure names commercial entities rather than state actors directly. The sustained economic incentive for such campaigns is clear regardless of attribution: as long as frontier model capabilities represent a significant competitive advantage, the cost-benefit calculus of distillation attacks favors continued investment by well-resourced actors.

Recommendations

Immediate Actions

For AI developers and organizations that provide proprietary AI capabilities through external APIs, the Anthropic incident provides concrete operational guidance on detection methodology. Cross-account behavioral correlation—identifying coordinated activity patterns across account clusters rather than monitoring individual accounts in isolation—is the detection approach that Anthropic found effective and that standard API gateway tooling does not perform by default. Organizations should evaluate whether their existing API monitoring infrastructure captures the metadata required for cross-account correlation (payment method patterns, infrastructure overlaps, prompt structure similarity, capability targeting patterns) and, if not, treat this as a gap that requires investment proportional to the competitive value of the underlying model.

Account verification programs that grant elevated access to API resources—researcher programs, education programs, startup programs—warrant immediate review in light of the fraudulent account cluster technique. If the primary defense against geographic and usage restrictions is account-level attestation, and if commercial services exist specifically to provide false attestation at scale, then the verification program's assurance value needs to be reassessed. Harder verification requirements for high-volume or high-risk account categories, including documentation requirements and manual review thresholds, reduce but do not eliminate this vector.

Short-Term Mitigations

Output watermarking represents an increasingly viable defense-in-depth measure that organizations developing proprietary AI capabilities should evaluate for deployment. Research frameworks including ModelShield [9] and MEA-Defender [10] embed cryptographically verifiable signals into model outputs that survive distillation training, enabling attribution of extracted model outputs back to their source. Watermarking does not prevent extraction, but it provides forensic evidence of the source of capabilities in extracted models and creates legal leverage for enforcement actions. Research indicates that character-level perturbations—typos, homoglyphs, spacing variations—may disrupt token-level watermarks, so organizations evaluating watermarking solutions should assess robustness against this evasion technique before deployment.

Rate limiting and output perturbation strategies that degrade extraction fidelity without materially harming legitimate use remain valuable as friction mechanisms, even if they cannot prevent a sufficiently motivated and well-resourced actor from achieving extraction over time. The goal is not to make extraction impossible—academic consensus holds that complete prevention is not achievable against a capable adversary [13]—but to make it prohibitively expensive relative to alternatives, and to ensure that the cost and time required create detection opportunities. Organizations should calibrate these controls to the competitive sensitivity of the capabilities being protected.

Industry intelligence sharing between AI providers, as practiced by Anthropic and Google GTIG in their coordinated February 2026 disclosures, significantly accelerates detection of threat actors who operate across multiple platforms. Security teams at organizations providing AI APIs should evaluate formal threat intelligence sharing arrangements with peer providers, particularly those that could enable correlated detection of hydra cluster infrastructure and shared proxy intermediaries.

Strategic Considerations

The intersection of model extraction and export control policy has implications for how organizations structure their API access programs. If regulatory regimes increasingly treat proprietary AI capabilities as controlled items—analogous to controlled dual-use technology under existing export regulations—organizations that provide commercial API access may face affirmative compliance obligations to implement detection and prevention controls against extraction by restricted parties, in addition to the contractual and economic incentives they already have. Security and legal teams should monitor this regulatory space actively, particularly as the OWASP GenAI Top 10 framework's classification of model theft under LLM10 gains traction as a compliance reference [11].

For organizations that depend on third-party AI APIs rather than developing proprietary models, the strategic implication runs in a different direction: the AI capabilities embedded in products and workflows may be less proprietary, and therefore less competitively differentiated, than currently assumed. As distillation-based capability transfer matures, the moat protecting frontier model capabilities will increasingly depend on continuous innovation rather than the durability of any particular model's behavioral profile. Organizations should factor this into AI vendor assessments and technology roadmaps.

CSA Resource Alignment

This research note connects to several existing CSA frameworks and initiatives that provide actionable guidance for organizations evaluating their exposure to model extraction threats.

The **CSA AI Controls Matrix (AICM)** provides a structured framework for evaluating AI-specific security controls across 18 domains, including AI supply chain security and access governance. The model extraction threat directly implicates AICM controls related to API access management, behavioral monitoring, and intellectual property protection. Organizations should reference the AICM's AI supply chain security domain when scoping controls for third-party AI API access programs [12].

The **CSA MAESTRO framework** for agentic AI threat modeling provides a structured methodology for identifying how AI systems can be manipulated through their interfaces. The hydra cluster attack pattern—in which individually normal-appearing interactions aggregate into an adversarial behavioral pattern—is precisely the kind of emergent threat that MAESTRO's layered threat analysis is designed to surface. Organizations using MAESTRO for threat model reviews of AI API services should explicitly include distillation attack patterns in their MAESTRO Layer 1 (AI model surface) analysis.

The **CSA Cloud Controls Matrix (CCM)** contains controls relevant to API security, identity and access management, and data classification that map to the infrastructure dimensions of model extraction risk. CCM domains covering application and interface security (AIS), identity and access management (IAM), and supply chain management (STA) provide the cloud-layer control structure into which AI-specific extraction defenses should be integrated. The CCM's supply chain domain is particularly relevant given the role of commercial proxy reseller services in enabling large-scale fraudulent account creation.

The **CSA Zero Trust guidance** is directly applicable to the detection gap identified in Anthropic's investigation: the principle that authenticated access does not imply trusted behavior should be operationalized in API monitoring architectures. A Zero Trust approach to AI API traffic would treat all traffic as potentially adversarial in terms of intent, even where authentication credentials are valid, and would focus monitoring on behavioral signals rather than access-layer signals. This reorientation is precisely what effective distillation attack detection requires.

References

[1] Anthropic, "Detecting and Preventing Distillation Attacks," Anthropic News, February 23, 2026. https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks

[2] P. Garraghan (Mindgard), "Model Leeching: An Extraction Attack Targeting LLMs," Mindgard Research, January 16, 2025. https://mindgard.ai/resources/model-leeching-an-extraction-attack-targeting-llms

[3] B. Jensen, "Protecting Our Edge: Trade Secrets and the Global AI Arms Race," CSIS, 2026. https://www.csis.org/analysis/protecting-our-edge-trade-secrets-and-global-ai-arms-race

[4] U.S. Copyright Office, "Copyright and Artificial Intelligence: Part 2 – Copyrightability," January 29, 2025. https://www.copyright.gov/ai/

[5] Google Threat Intelligence Group, "Distillation, Experimentation, and Integration: AI Adversarial Use," Google Cloud Blog, February 2026. https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use

[6] Treblle, "The Attack That Looked Like Nothing at All: Anthropic's Distillation Breach Breakdown," Treblle Blog, 2026. https://treblle.com/blog/anthropic-distillation-breach-breakdown

[7] Z. Liang et al., "Yes, My LoRD: Guiding Language Model Extraction with Locality Reinforced Distillation," arXiv:2409.02718, September 4, 2024. https://arxiv.org/abs/2409.02718

[8] Sysdig Threat Research Team, "LLMjacking: Stolen Cloud Credentials Used in New AI Attack," Sysdig Blog, May 2024. https://www.sysdig.com/blog/llmjacking-stolen-cloud-credentials-used-in-new-ai-attack

[9] K. Pang et al., "ModelShield: Adaptive and Robust Watermark against Model Extraction Attack," arXiv:2405.02365, 2024. https://arxiv.org/abs/2405.02365

[10] P. Lv et al., "MEA-Defender: A Robust Watermark against Model Extraction Attack," arXiv:2401.15239, 2024. https://arxiv.org/abs/2401.15239

[11] OWASP, "LLM10: Unbounded Consumption," OWASP GenAI Top 10 2025. https://genai.owasp.org/

[12] Cloud Security Alliance, "AI Controls Matrix (AICM): Introductory Guidance," CSA, 2025. https://cloudsecurityalliance.org/artifacts/ai-controls-matrix

[13] K. Zhao et al., "A Survey on Model Extraction Attacks and Defenses for Large Language Models," arXiv:2506.22521 / ACM SIGKDD 2025, June 2025. https://arxiv.org/abs/2506.22521

[14] Winston & Strawn LLP, "Is AI Distillation By DeepSeek IP Theft?," Winston & Strawn Insights, 2026. https://www.winston.com/en/insights-news/is-ai-distillation-by-deepseek-ip-theft