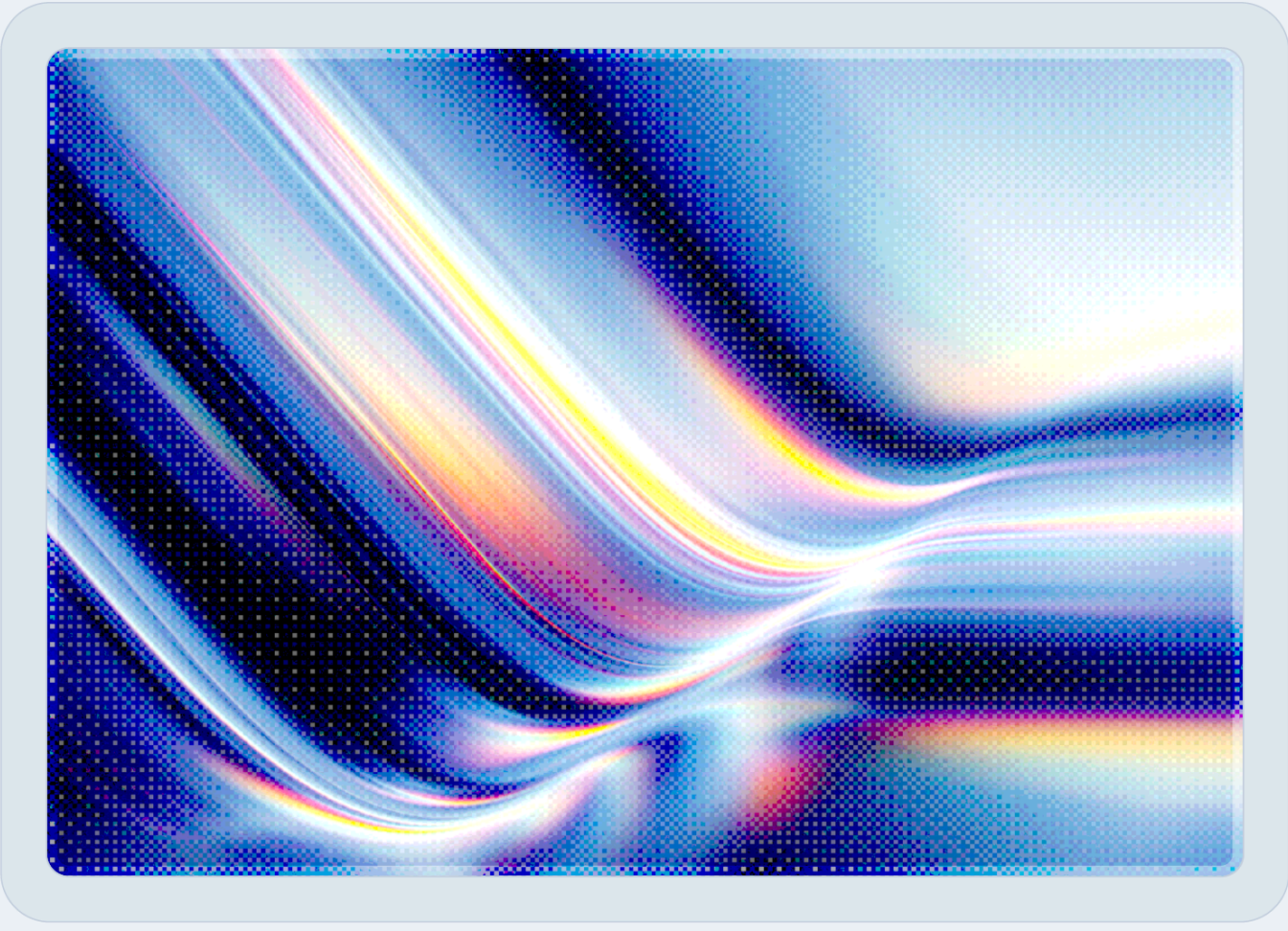
CSAI Foundation | Cloud Security Alliance

Foundation Model IP Theft: Threat Model for AI Labs

How State-Sponsored Actors and Competitive Espionage Are Targeting Model Weights, ML Infrastructure, and AI Supply Chains

2026-05-17

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Foundation models represent a new class of high-value intellectual property: training a frontier model can require hundreds of millions of dollars in compute and years of research, making model weights a target of comparable strategic interest to trade secrets in pharmaceuticals or semiconductor design.
- Nation-state threat actors — particularly those attributed to the People's Republic of China — have shifted from broadly targeting AI company infrastructure to specifically pursuing model weights, training data, and the algorithmic techniques that differentiate frontier models from their commercial derivatives [1].
- Model extraction via API-based distillation represents an attack path that requires no infrastructure breach, making it accessible to a broader range of adversaries than traditional intrusion-based IP theft: adversaries can systematically query production model APIs to generate training data for surrogate models without ever compromising the target's network perimeter [2].
- The ML operations toolchain — including experiment tracking platforms (MLflow), model registries, GPU orchestration layers, and open-source model hubs (Hugging Face) — introduces a broad attack surface that is inadequately hardened against adversarial access, with documented critical vulnerabilities enabling remote code execution and model exfiltration [3][4].
- The January 2025 export control rule creating ECCN 4E091 represents the first formal U.S. government recognition that frontier model weights constitute controlled technology, though its specific thresholds and licensing framework have been subject to revision by subsequent executive action [5].
- Organizations operating, deploying, or integrating foundation models must treat model weights with the same classification rigor and access controls applied to their most sensitive proprietary source code, recognizing that a single successful exfiltration event may transfer the equivalent of years of research investment and hundreds of millions of dollars in compute cost to an adversary who incurred none of that development expenditure [1].

Background

The competitive dynamics of foundation model development have produced an unusual alignment between commercial rivalry and geopolitical competition. A small number of organizations — primarily headquartered in the United States and China — are building the world's most capable AI systems. The economic moats these systems create are substantial: the largest frontier models incorporate years of research, billions of dollars in compute expenditure, and the cumulative labor of thousands of engineers and researchers. Unlike most software intellectual property, a foundation model's value is concentrated in a relatively small set of files. The model weights, which encode all of the learned knowledge and behavioral patterns of the system, can be copied in a matter of hours over high-bandwidth connections and replicated at marginal cost by any adversary who obtains them.

This concentration of value in a portable, copyable artifact creates a threat profile unlike most enterprise software targets. Historically, IP theft from a technology company might yield source code that still required engineering effort to adapt, deploy, and maintain. Stolen model weights, by contrast, can be loaded into compatible inference infrastructure without any further training investment — there is no source code to compile, no additional engineering to adapt, and no research work to reproduce. An adversary who has the requisite GPU infrastructure can begin serving the stolen model within days, rather than the years required to train an equivalent system independently. That adversary obtains not just the system's current capabilities but the research insights, training methodology, and alignment techniques that went into producing it — knowledge that may take years for a competitor working independently to replicate.

The RAND Corporation's 2024 analysis of model weight security identified this dynamic explicitly, cataloging the potential adversary population to include nation-states seeking strategic advantage, commercial competitors seeking to eliminate R&D cost disadvantages, and criminal actors seeking to monetize model capabilities [1]. Subsequent threat intelligence has corroborated these threat actor profiles: the CrowdStrike 2026 Global Threat Report documents continued acceleration of state-sponsored campaigns targeting AI organizations, with specific focus on model weights, training methodologies, and proprietary training data [9]. RAND's analysis further emphasized that the threat extends beyond traditional espionage to include scenarios where stolen weights are weaponized — for instance, deploying a model without the safety training and alignment work the original developer invested in, or fine-tuning a stolen model for offensive applications without constraints.

The geopolitical dimension of this threat crystallized publicly in early 2025 when reports emerged that Chinese AI companies had conducted systematic API distillation campaigns against frontier model providers, including attempts to extract training data and behavioral patterns from Claude, GPT-4, and other leading models [2]. These campaigns, which according to contemporaneous reporting involved

creating tens of thousands of accounts to query models at scale, represented a new variant of competitive intelligence gathering that exploited the same open-access interfaces that make commercial AI products broadly useful.

Security Analysis

The Attack Surface Specific to AI Labs

AI model providers face an attack surface that combines the general enterprise threat landscape with several categories of risk unique to the ML development lifecycle. Understanding this expanded surface is prerequisite to effective defense.

Among the most sensitive assets in this environment are the model weights themselves. Frontier models may occupy hundreds of gigabytes of storage when represented in full precision, and must be distributed across large numbers of GPUs during both training and inference. During training, weights are continuously updated and checkpointed; during inference, they reside in GPU high-bandwidth memory (HBM) where they are directly accessible to any process with sufficient privilege on the host or the virtualization layer beneath it. A privileged attacker — whether a cloud provider insider, a compromised orchestration platform, or an adversary who has obtained root access through a vulnerability in the GPU software stack — can read model weights directly from GPU memory during inference, bypassing any application-layer access controls.

Beyond the weights themselves, the ML development lifecycle involves multiple additional data classes that carry significant strategic value. Training datasets, curated from web crawls, licensed sources, and proprietary human feedback, require substantial investment to assemble and represent years of iterative refinement. Fine-tuning datasets, particularly those used for instruction following, safety training, and alignment, encode research insights that are difficult to reconstruct. Evaluation benchmarks, hyperparameter configurations, and ablation study results collectively constitute a research record that an adversary could use to substantially accelerate independent development.

Infrastructure Vulnerabilities in the ML Toolchain

The tooling ecosystem that supports modern ML development was designed primarily around research productivity goals — rapid iteration, experiment sharing, and training scalability — rather than adversarial security. Platforms like MLflow, Kubeflow, and Ray were built to help teams manage experiments, share models, and scale training jobs, not to protect those assets from sophisticated adversaries. Several documented vulnerabilities illustrate the risk this creates.

MLflow, the most widely adopted open-source experiment tracking platform, has been the subject of multiple critical vulnerability disclosures. Researchers identified attack paths enabling remote code execution, arbitrary file write, and local file include attacks against exposed MLflow instances [3]. A particularly significant attack variant allows adversaries who have obtained limited access to an MLflow environment to redirect model artifact storage to attacker-controlled cloud buckets, causing new training runs to automatically exfiltrate their outputs to external infrastructure. Researchers assessing the attack vector noted that cloud object storage is a common backend configuration for production MLflow deployments, making this path broadly applicable [3].

The Hugging Face model hub, which hosts hundreds of thousands of open-source models and serves as a distribution point for many commercial and research projects, has emerged as a critical AI supply chain vector. Researchers have documented numerous malicious model repositories on the platform, many exploiting the fact that the most widely used model serialization format — Python's `pickle` — inherently supports arbitrary code execution during deserialization [4]. A model file using `pickle` format is effectively an executable: loading it runs whatever code the creator embedded, with no reliable static analysis tool capable of fully enumerating what that code will do. CVE-2025-1716 specifically documents a bypass technique against `picklescan`, one of the primary defensive tools deployed to detect malicious `pickle` payloads in model files [13]. The security implication for organizations that load open-source models into their infrastructure is that model files must be treated with the same rigor applied to third-party executables — scanned, sandboxed, and verified against cryptographic hashes from trusted sources.

NVIDIA's GPU software stack has accumulated an increasing volume of security vulnerabilities as the stack has grown in complexity. The NVIDIA Container Toolkit and GPU Operator, which mediate GPU access for containerized workloads, have received patches for multiple vulnerabilities including privilege escalation paths that could allow a workload with standard GPU access to obtain elevated privileges on the host system — CVE-2024-0132, a container breakout vulnerability disclosed in late 2024, is a representative example [6]. For AI providers running multi-tenant GPU infrastructure, these vulnerabilities represent potential paths for tenant-to-tenant or tenant-to-host attacks that could expose model weights or training data belonging to other customers.

API-Based Model Extraction and Distillation

The lowest-sophistication attack against a model provider does not require breaching any infrastructure. Model extraction via systematic API querying — commonly called distillation attack or model stealing — allows an adversary to train a surrogate model that approximates the target's behavior by studying its

input-output relationships. The technique exploits the fact that every commercial model inference API reveals information about the model's output distribution through the patterns of its responses, providing a rich supervisory signal for training a surrogate.

Contemporaneous reporting on Chinese AI lab activity in early 2025 described a structured application of this approach: adversaries created tens of thousands of accounts to query Claude at scale, constructing large-scale datasets of model responses designed to maximize coverage of the model's behavioral space [2]. Similar reports emerged regarding other frontier model providers. The distillation technique is particularly effective for capturing a model's reasoning style, formatting conventions, and high-level capabilities, even if it cannot reproduce the exact weight values. For a competitor willing to invest modest resources in API access, the cost-benefit calculus is stark: the alternative is spending hundreds of millions of dollars to train an equivalent model from scratch.

Detection of distillation campaigns presents genuine challenges. The individual API calls that constitute a distillation campaign are often indistinguishable from legitimate high-volume usage when examined in isolation — the distinguishing signals emerge at the population level. Detection depends on behavioral heuristics: unusually systematic coverage of the input space, high-entropy prompting patterns designed to probe model boundaries, correlated activity across accounts sharing infrastructure signatures, or implausibly high query volumes relative to any plausible business use case. These signals require behavioral analytics across the full population of API users — a capability that most model providers have developed to varying degrees but that remains an active area of investment.

Insider Threat and Social Engineering

The insider threat surface at AI labs is elevated relative to most enterprise environments because of the concentration of highly privileged access required for model development work. Training engineers, safety researchers, and infrastructure operators routinely have access to model checkpoints, training data, and inference infrastructure that represent the organization's most sensitive IP. The combination of competitive talent markets, high compensation expectations, and interest from foreign intelligence services creates conditions under which insider risk warrants systematic program investment.

Social engineering attacks targeting AI lab employees have adapted to AI tooling itself. AI-generated spear phishing messages, synthesized voice samples of leadership figures, and deepfake video content have all been documented in attacks against technology company employees [7]. The increasing specificity achievable with AI-generated social engineering — drawing on publicly available information about the target's role, projects, and professional relationships — likely reduces the effectiveness of traditional awareness training, which was designed for more generic phishing templates; Microsoft's April 2026 threat analysis found that AI-enhanced phishing achieved click-through rates significantly higher

than conventional approaches [10]. AI labs should assume that their employees are among the highest-priority social engineering targets for sophisticated adversaries and structure their security awareness programs accordingly.

Geopolitical and Regulatory Context

The U.S. government's recognition that frontier model weights constitute controlled technology marks a significant evolution in how AI IP is understood from a national security perspective. The Bureau of Industry and Security's January 2025 rule creating ECCN 4E091 specifically addresses model weights for AI systems that exceed defined computational thresholds [5]. While the rule's specific thresholds and licensing framework have been subject to revision by subsequent executive action, the underlying principle — that model weights for sufficiently capable systems warrant the same export control treatment as other dual-use technologies — reflects a growing recognition, shared across administrations, of the strategic importance of frontier AI.

Export controls create compliance obligations but also provide a legal framework for characterizing certain forms of model theft as violations of arms control statutes rather than civil IP infringement alone. For organizations operating under these controls, the obligation to track where model weights travel, who can access them, and under what licensing conditions they can be shared creates documentation requirements that, when implemented effectively, also serve as a forensic record for incident response.

Recommendations

Immediate Actions

Organizations that develop, operate, or integrate frontier AI models should take several near-term actions to reduce exposure to the most acute risks described in this note.

Model weight inventories are a prerequisite for effective protection. Organizations should identify all locations where trained model checkpoints and production weights are stored, including primary storage, backups, experiment tracking systems, and staging environments. Every location should have a documented access control policy specifying which personnel and systems may read, copy, or move model files, and access should be logged with sufficient fidelity to support forensic analysis in the event of a suspected exfiltration.

ML operations tooling should be patched promptly and exposed only on networks with strict access controls. MLflow instances and equivalent experiment tracking systems should not be exposed to the public internet. Model registry interfaces should require authentication and ideally operate on isolated network segments. The principle of least privilege applies with particular force here: researchers should have access to their own experiment artifacts and to shared model repositories as needed for collaboration, but not global read access to all model artifacts in the organization.

Model files sourced from open registries, including Hugging Face and similar platforms, should be scanned with serialization-aware security tools before being loaded into any production or development environment. Organizations should establish an internal model registry where approved and vetted external models are hosted after passing security review, rather than allowing direct loading from external sources in production pipelines. Cryptographic verification of model file hashes against values published by the original author provides an additional layer of assurance.

Short-Term Mitigations

API access policies for commercial model interfaces should incorporate behavioral analytics to detect distillation campaigns. Rate limiting alone is unlikely to be sufficient, as a determined adversary can pace queries to stay within nominal thresholds while still systematically covering the model's behavioral space; effective detection requires modeling the behavioral fingerprint of legitimate use and flagging deviations consistent with systematic capability extraction. Organizations providing model APIs should invest in behavioral anomaly detection and should establish clear policies about terms of service violations that constitute extraction attempts.

Access to model weights in production inference infrastructure should be protected through a defense-in-depth architecture that addresses both the application layer and the infrastructure layer. Confidential computing approaches, including trusted execution environments (TEEs) and emerging confidential GPU capabilities from NVIDIA and AMD, can protect model weights from privileged-access attacks at the infrastructure level by maintaining encryption of model data even in GPU memory [8]. While these technologies carry deployment complexity and performance overhead, they represent a meaningful control against adversaries who have achieved infrastructure-level access.

Insider threat programs at AI organizations should be calibrated to the elevated sensitivity of model IP. This includes periodic access reviews to ensure that research personnel who have changed roles or left the organization no longer retain access to model artifacts, behavioral monitoring of access patterns to high-sensitivity model repositories, and clear policies about the handling of model checkpoints on personal devices or external systems.

Strategic Considerations

The long-term security posture of AI model providers depends on treating model IP protection as a program-level investment rather than a checklist of individual controls. Several dimensions merit sustained attention.

The supply chain through which training data, open-source components, and pre-trained base models flow into frontier model development creates upstream dependencies that carry their own risk. An adversary who can influence what a model is trained on — through data poisoning in web crawls, compromised pre-training checkpoints, or manipulated fine-tuning datasets — may achieve effects more durable than simple weight exfiltration. Supply chain assurance for AI training data and model components should become a standard part of model development governance.

Watermarking and fingerprinting techniques for model weights are an active area of research. Methods that embed verifiable, statistically detectable markers into model weights with minimal performance impact could provide both a forensic mechanism for attributing stolen models and a deterrent against certain distillation techniques. Organizations should monitor this research area and evaluate emerging solutions as they mature.

The policy and legal landscape for AI IP protection is evolving rapidly. Export control regimes, intellectual property frameworks, and emerging AI-specific legislation all have implications for how model theft is characterized and prosecuted. Organizations should engage proactively with policymakers and industry bodies to ensure that legal frameworks keep pace with the technical reality of the threat, and should maintain relationships with law enforcement and intelligence community contacts who can support incident response when state-sponsored actors are involved.

CSA Resource Alignment

This research note connects to several existing Cloud Security Alliance frameworks and initiatives.

The AI Controls Matrix (AICM) provides the most directly applicable governance structure for AI model providers facing the risks described in this note [11]. The AICM's Model Provider (MP) domain addresses security controls spanning model training, deployment, and supply chain integrity. The AICM Implementation Guidelines for Model Providers specifically address data protection, access control, and supply chain security obligations that correspond to the attack vectors documented here. Organizations should assess their coverage of AICM Model Provider controls as a starting point for gap identification.

The MAESTRO threat modeling framework provides a structured methodology for modeling adversarial threats against AI systems at each layer of the AI stack [12]. For model providers, MAESTRO's analysis of infrastructure-layer threats, supply chain risks, and insider vectors maps directly to the categories described in the Security Analysis section of this note. MAESTRO scenarios involving model weight exfiltration and training data compromise provide ready-made threat models that security teams can adapt to their specific environments.

The CSA Zero Trust Guidance is particularly relevant for the ML operations environment. The principle of never implicitly trusting any request for access to model artifacts — regardless of network location, identity, or prior access history — addresses the lateral movement risk that makes infrastructure-level compromises of AI lab environments so consequential. Zero Trust architectures that enforce continuous verification for access to model registries and training checkpoints reduce the blast radius of any individual credential or system compromise.

The STAR (Security Trust Assurance and Risk) program offers a mechanism for AI organizations to demonstrate their security posture to customers and partners. As enterprise procurement increasingly requires AI providers to demonstrate security assurance, STAR-based attestation can serve both as a credibility marker and as a structured driver for internal security program maturity.

References

[1] RAND Corporation. "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models." RAND Corporation, 2024.

[2] VentureBeat. "Anthropic Says DeepSeek, Moonshot, and MiniMax Used Fake Accounts to Distill Claude." VentureBeat, February 2025.

[3] CSO Online. "MLflow Vulnerability Enables Remote Machine Learning Model Theft and Poisoning." CSO Online, December 2023.

[4] Kellas, A.D. et al. "PickleBall: Secure Deserialization of Pickle-based Machine Learning Models." arXiv, 2025.

[5] Sidley Austin LLP. "New U.S. Export Controls on Advanced Computing Items and Artificial Intelligence Model Weights." Sidley, January 2025.

[6] NIST NVD. "CVE-2024-0132 Detail." National Institute of Standards and Technology, 2024.

[7] SecureTrust. "Phishing in 2026: AI-Driven Attacks, Deepfakes, and the Next Wave of Cyber Threats." SecureTrust, December 2025.

[8] NVIDIA. "Confidential Computing on NVIDIA H100 GPUs." NVIDIA, 2024.

[9] CrowdStrike. "2026 CrowdStrike Global Threat Report." CrowdStrike, February 2026.

[10] Microsoft Security Blog. "Threat Actor Abuse of AI Accelerates from Tool to Cyberattack Surface." Microsoft, April 2026.

[11] Cloud Security Alliance. "AI Controls Matrix (AICM)." CSA, 2024.

[12] Cloud Security Alliance. "MAESTRO: AI Threat Modeling Framework." CSA, 2024.

[13] NIST NVD. "CVE-2025-1716 Detail." National Institute of Standards and Technology, February 2025.

