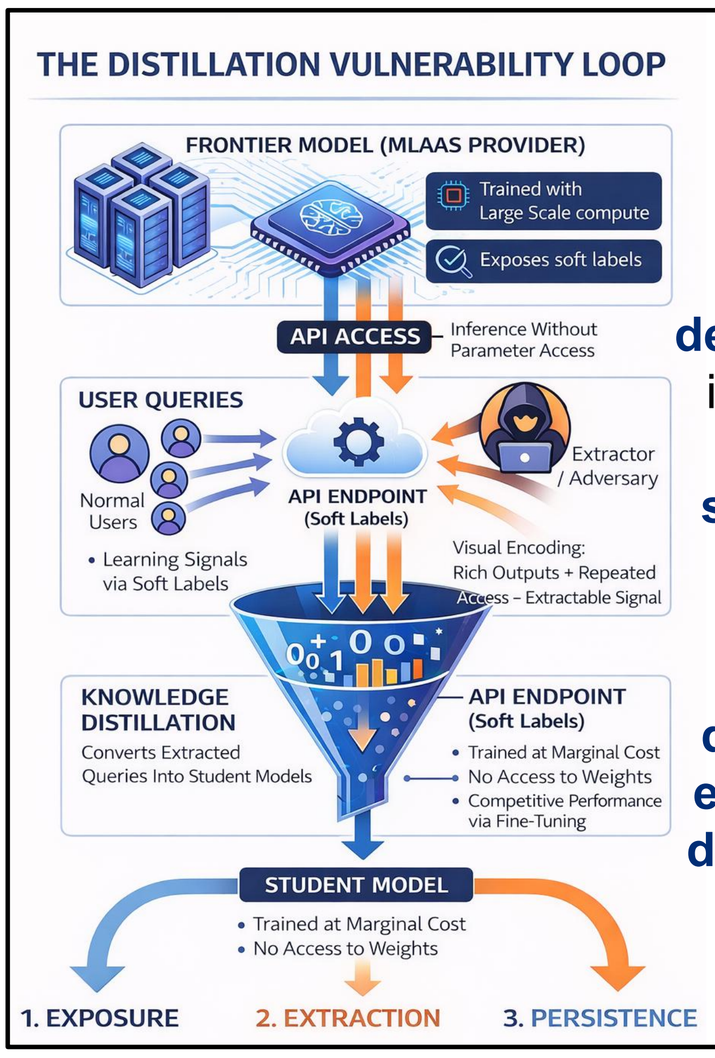
Securing Frontier AI / ML Models Against Extraction Attacks via Distillation Techniques Using An Offensive Deterrence Framework

Problem Statement and Goals



Goal:

A **techno-policy shift** toward an **offensive deterrence posture**, in which platforms not only **actively surface verifiable signals** that the model has been derived through **distillation-based extraction** but also **degrade or disrupt** distilled **student performance**.

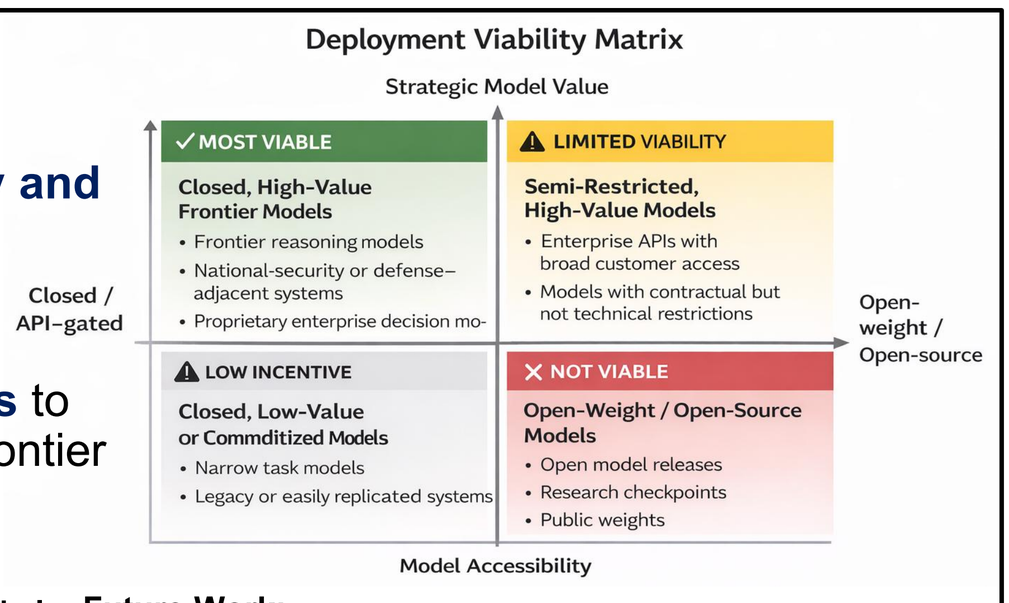
Approach

- Built ***Kepler***, an interactive **distillation attack simulator** that demonstrates extraction mechanics.
- Surveyed **extraction-mitigation policies** adopted by **frontier labs** + major **MLaaS platforms**.
- Studied **prominent extraction + distillation attempts** on frontier labs by **state-linked + private actors**.
- Identified **structural and technical gaps** that constrain **existing responses** to detection and disruption.
- Used *Kepler* to simulate a basic framework that combines **active watermarking** with **offensive characteristics**.
- Evaluated effectiveness on (a) **detectability + survival under distillation** (b) **student model degradation** and (c) **deployment considerations**.

Results

Key Results:

- Distilled students exhibit **measurable drops in task accuracy and calibration**, with **degradation compounding** under repeated distillation cycles.
- As **scaling gains plateau**, distillation enables **student models** to **remain competitive** through **recursive fine-tuning** despite frontier architectural updates.
- Extraction can **target specific capabilities** apart from full model replication, reinforcing the need for function-level defenses.
- Analysis highlights a **deterrence gap** driven by **platform reluctance** to deploy visibly **aggressive countermeasures**.
- **Selective viability** of the framework for closed, **high-value frontier models**, while **open-source** deployments remove the leverage required for deterrence.



Future Work:

- Establish legal and ethical boundaries that distinguish offensive watermarking from poisoning or discriminatory service behavior.
- Define robust decision thresholds and rollback mechanisms to limit misclassification risk and collateral degradation for benign users.
- Develop governance and disclosure strategies that enable selective, low-visibility deployment without accelerating adversarial countermeasures or extraction arms races.

Ziauddin Sherkar
Johns Hopkins University School of Advanced International Studies

Securing Frontier AI / ML Models Against Extraction Attacks via Distillation Techniques Using An Offensive Deterrence Framework

Abstract

Frontier AI models and modern Machine Learning as a Service (MLaaS) platforms face a core vulnerability paradox. The probabilistic outputs that enable high-quality downstream performance, including soft labels and confidence scores, also expand the attack surface for distillation-based extraction. This creates a structural asymmetry in which inference-level access allows adversaries to convert a provider's training compute into competitive student models at marginal cost, without access to model weights. Distillation exploits this asymmetry by transferring foundational capabilities through repeated querying and training rather than direct parameter access.

Historical precedents demonstrate the strategic effectiveness of this approach. Chinese military-civil fusion programs and private actors have repeatedly used distillation techniques to close capability gaps by leveraging frontier models and proprietary APIs to accelerate domestic model development. Contemporary examples include the extraction of reasoning trajectories and synthetic data from OpenAI reasoning models to refine the DeepSeek series. Commercial competitors and non-state actors have similarly adopted distillation due to its low cost and high leverage, allowing student models to approach competitive performance with significantly reduced computational investment.

As scaling gains plateau and benchmark improvements slow, model progress increasingly favors incremental refinement over generational leaps. In this regime, distilled student models are less likely to be rendered obsolete by architectural updates and can remain competitive through recursive fine-tuning and repeated distillation. Importantly, extraction often targets specific capabilities rather than full model replication, complicating traditional ownership-based defenses.

This work advocates a techno-policy shift from a strictly detection-defensive positioning toward an offensive deterrence posture. Using *Kepler*, an interactive distillation attack simulator, it analyzes existing mitigation policies, real-world extraction attempts, and structural gaps in detection-focused responses. Adaptive output perturbations combined with distillation-surviving watermark signals can degrade unauthorized student models without affecting frontier model performance. Findings reveal a deterrence gap driven by platform reluctance to deploy visibly aggressive countermeasures and indicate that such approaches are selectively viable for closed, high-value frontier deployments, while open-source releases remove the leverage required for effective deterrence.

Ziauddin Sherkar
Johns Hopkins University School of Advanced International Studies

JOHNS HOPKINS SCHOOL of ADVANCED INTERNATIONAL STUDIES