ISSUE BRIEF

FRONTIER MODEL FORUM

Issue Brief: Adversarial Distillation

DATE

February 23, 2026

Introduction

Adversarial distillation is a growing concern for frontier AI safety and security. Although frontier AI developers may use or authorize distillation in select cases – for instance, to create smaller, more efficient, versions of their flagship models or to adapt models for specialized domains, such as life sciences – adversarial distillation may raise significant safety and security risks by allowing malicious actors to replicate core model capabilities without necessarily replicating the accompanying safety measures.

This issue brief aims to raise awareness of adversarial distillation techniques without discouraging legitimate uses of frontier AI models, including authorized distillation. The dual use nature of distillation makes this balance particularly challenging. By providing an overview of common adversarial distillation methods and target categories, we aim to help the AI ecosystem develop shared norms around responsible model use and inform the development of effective technical and policy safeguards.

Authorized Distillation and Adversarial Distillation

Distillation is a process by which a secondary model learns to replicate the knowledge and capabilities of an originator model by training on its outputs. It is leveraged in the AI ecosystem as a cost-effective and computationally efficient way to transfer skills and capabilities from one model to another model. Put differently, this technique enables efficient knowledge and capability transfer from large, computationally expensive "teacher" models to smaller "student" models with minimal performance loss.

Authorized distillation involves distillation within a company or distillation that occurs where there is a permissive license with an open source model. Authorized distillation employs different techniques and different levels of access to the teacher models' internal probability distribution or hidden states. Common approaches include training the student model on the teacher model's inner activations or training on the teacher model's inputs and outputs.

Authorized distillation has been instrumental in driving innovation of large language models as well as enabling deployment in scenarios requiring task-specific optimization and runtime efficiency. Authorized distillation has several advantages. It allows companies and researchers to compress versions of large models into small language models that can deliver on-device AI, limit inference cost, reduce latency, and optimize models for specific tasks.

In contrast, adversarial distillation often involves covert or indirect access where malicious actors only have access to the teacher model's outputs, which they use to train a student model, and thus replicate the teacher model's capabilities. Often, adversarial distillation involves violating a model developer's terms of use or license to enable training on the teacher model's outputs. The resulting capabilities of the student model are determined, in part, by how many outputs malicious actors can obtain from the teacher model, as well as the amount of compute available to train the new model. When the core capabilities of frontier AI models are replicated by malicious actors without replicating the corresponding safety measures, this may lead to the use of powerful AI models without appropriate safety measures applied, increasing the risk of misuse and harmful applications.

Adversarial Distillation Method Taxonomy

One of the most urgent adversarial distillation approaches involves Chain-of-Thought (CoT) reasoning, or the intermediate natural language reasoning steps that a large language model articulates before arriving at its final answer. Since CoT traces reflect how frontier AI models reason their way to a final output, eliciting and critiquing CoT outputs are a prime target for adversarial distillation. Although companies powering AI systems with reasoning models typically obscure their models' full CoT reasoning from users, adversarial actors work to extract it. Once collected, malicious actors can use reasoning sequences from powerful models as training data to fine-tune secondary models.

The below constitute common adversarial CoT distillation methods. However, each of the methods discussed can be used for benign reasons, making adversarial distillation detection a significant technical challenge:

- **Chain-of-Thought Exfiltration**: Malicious actors design prompts that include a scenario and requirements for how the model should answer. They may instruct a model to provide Chain-of-Thought reasoning when solving difficult STEM problems or answering complex multimodal questions. For example, a

malicious actor might generate <prompt, answer, CoT> triplets based on multimodal input, such as an image or pdf. By collecting these reasoning traces at scale, malicious actors build training datasets for their own models.

- **Chain-of-Thought Critiquing**: Malicious actors provide candidate Chain-of-Thought sequences and ask the target model to critique and improve them. Through the feedback, the malicious actors can infer details about how the target model structures its reasoning, effectively reverse-engineering its problem-solving approach.

- **Chain-of-Thought Autograding**: As AI development has accelerated, researchers have increasingly automated the evaluation process by using large language models to grade outputs from other models. These large language model evaluators are known as autograders. Malicious actors can exploit a powerful model's autograding capabilities by prompting it to explain its complete evaluation methodology. Rather than simply receiving a score, malicious actors extract the detailed reasoning the model uses to assess responses, including its criteria for quality, how it identifies errors, and how it weighs different factors in its judgment. This grading chain-of-thought reveals sophisticated reasoning capabilities that can be transferred to secondary models.

In addition to CoT distillation, there are also several other common adversarial distillation methods. As above, these same methods may be used for benign reasons:

- **Prompt Generation for Reinforcement Learning**: Malicious actors develop specific prompts that are fed to both a student and teacher model to generate responses, and to train a discriminator to distinguish between teacher and student outputs. The student model is then optimized to produce responses that the discriminator cannot distinguish from those of the teacher. Through this adversarial training process, the student model receives implicit feedback on how its outputs compare to the teacher model's generations. Over time, this feedback loop enables the student model to progressively mimic the capabilities of the teacher model.

- **Synthetic Data Generation**: Malicious actors generate synthetic training data by feeding carefully selected prompts into a powerful model and collecting its responses. This creates a dataset of input-output pairs that captures the target model's capabilities. The secondary model can then be fine-tuned on this synthetic dataset, bypassing the need for manually curated training data.

Common Target Categories

Malicious actors target different capabilities when attempting to distill models, including:

- **Mathematical or scientific reasoning**. Capabilities that enable models to solve complex quantitative problems, derive proofs, or conduct multi-step

scientific analysis. These are core analytical competencies used in STEM tasks, which can be transferred through distillation and may have an impact on CBRN-relevant scientific domains.

- **Agentic coding and tool use**. Advanced skills that allow models to write, debug, and autonomously execute code across extended workflows. These are competencies that enable models to coordinate actions across tools or environments, making them highly valuable targets for malicious actors seeking to construct agentic or semi-autonomous systems. In this case, malicious actors seek to transfer not only reasoning capability but full-task solving behavior from large language model-based agents. Distilling advanced coding capabilities could increase cybersecurity risk, as distilled models retain strong coding abilities while having learned little, if any, of the safety training of their source models.

- **Multimodal processing**. Capabilities that enable a model to interpret, reason over, and generate insights from images, diagrams, or other non-text modalities. Malicious actors exploit these capabilities by submitting visual or multimedia files to infer training data. Common techniques include requesting labels for visual content to build training datasets, or asking the model to generate and then analyze visualizations, such as graphs or diagrams. Cross-modal reasoning abilities—such as parsing diagrams, reading scientific figures, or extracting structured data from images— could enable malicious actors to reconstruct sensitive workflows or identify vulnerabilities when transferred to uncontrolled models.

- **General reasoning**. Broad problem-solving capabilities that support coherent planning, explanation, and decision-making. Since general reasoning skills enable models to approach unfamiliar tasks with high proficiency, they are particularly attractive targets for malicious actors seeking to build systems capable of carrying out or assisting attacks across diverse environments.

Note that the choice of capability may depend on a malicious actor's intended use case for the distilled model. For example, a malicious actor seeking to create a model with strong technical reasoning might focus on extracting mathematical and scientific capabilities by prompting the frontier model with challenging STEM questions and using the CoT outputs to train their model. By contrast, a malicious actor planning offensive cyber operations might focus on distilling a model's advanced coding capabilities.

As the descriptions note, selectively extracting high-leverage capabilities via adversarial distillation introduces significant safety and security risks, since those capabilities typically require strict safeguards to deploy responsibly.

Conclusion and Takeaways

Adversarial distillation, particularly for the purpose of achieving malicious objectives, represents a growing concern for AI safety and security, requiring careful attention from developers and the broader AI community. While distillation itself is a valuable technique with numerous legitimate applications, adversarial distillation poses distinct risks.

The challenge lies in addressing these security concerns without impeding legitimate research and development activities. As this issue brief has outlined, adversarial distillation varies in method and targets different capability categories, with success depending on factors including access to model outputs and available computational resources. Understanding these methods and target categories is essential for developing effective countermeasures.

By documenting the primary methods and targets of adversarial distillation, we aim to raise public awareness of the issue and foster greater capacity across the frontier AI ecosystem to identify the risks posed by adversarial distillation.