



White Paper

Will AI distillation undermine public access to frontier models?

Content

01 Will AI distillation undermine public access to frontier models?

Pages 03-04

02 Why high investment is necessary

Pages 05-09

- Scaling laws: The exponential cost of intelligence
- Shifting the curve: Making AI more efficient but not cheaper
- Shifting the paradigm: A new frontier in AI training

03 Regulatory protections: A failing safeguard?

Pages 10-12

- The dilemma of data provenance and AI ethics

04 How can developers of frontier models protect their investment?

Pages 13-15

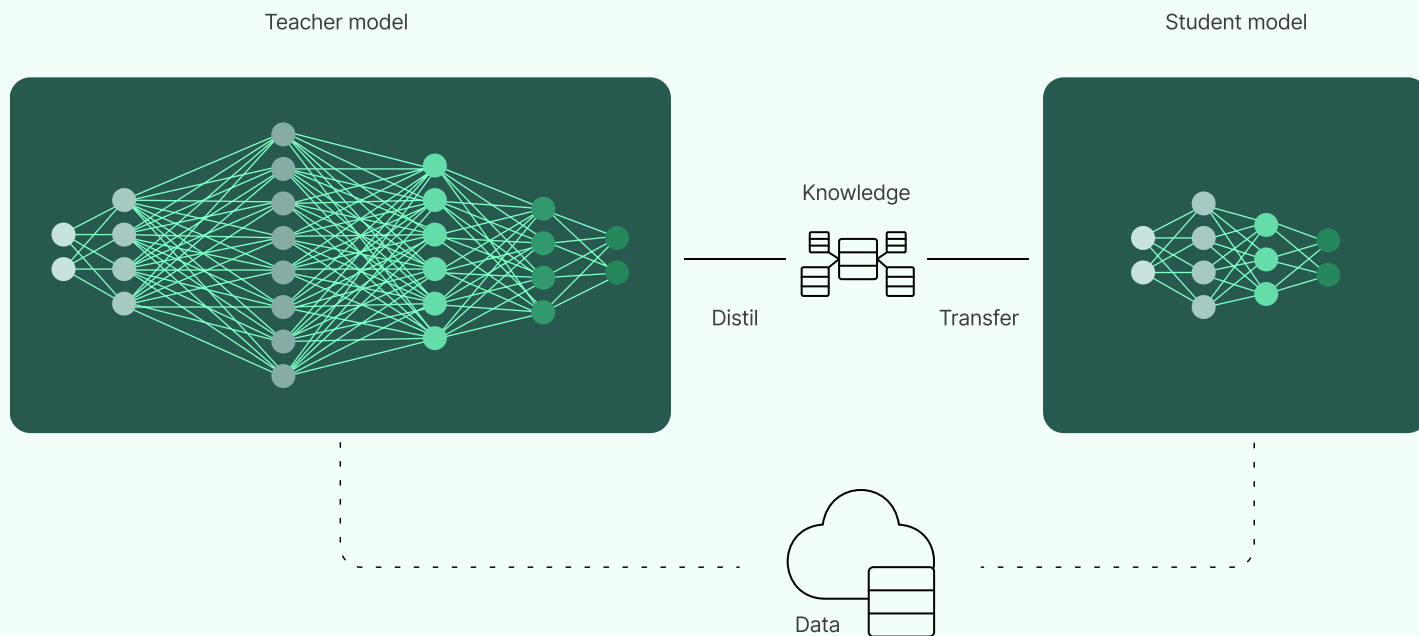
- Expanding protections against model distillation

05 The Future of AI and Public Benefit

Pages 16

* Author:

MILOŠ CIGOJ | Director of Healthcare and Regulatory Quality and Life Sciences at HTEC



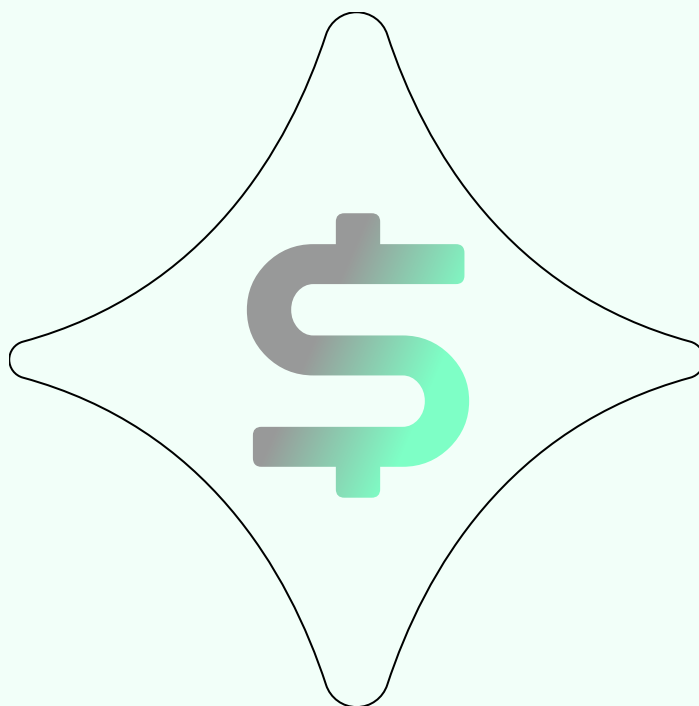
The rapid pace of artificial intelligence development is generating controversy due to the emergence of unauthorized distillation of commercial “teacher” models. Does this new phenomenon stifle further frontier research by undermining financial incentives and introducing ethical dangers?

On the one hand, leading AI labs such as OpenAI and Anthropic invest hundreds of millions into proprietary teacher models, providing public API access to drive breakthroughs in medical care, life sciences, engineering, education, and language translation. These frontier models are becoming increasingly expensive to train: OpenAI’s GPT-4 reportedly consumed \$78 million in compute, while Google’s upcoming Gemini Ultra required an estimated \$191 million in infrastructure. Meanwhile, major players—including OpenAI, Anthropic, Hugging Face, and Inflection—have secured massive funding rounds, underscoring the global AI arms race to develop the next wave of generative models.

At the same time, student models—derived via knowledge distillation from these expensive teacher models—can, in some cases, be open-sourced, allowing a broader pool of developers to leverage near-state-of-the-art performance at a fraction of the cost. This raises critical concerns about financial sustainability: If the massive R&D investment behind frontier models can be bypassed through distillation, will AI frontier research remain economically viable?

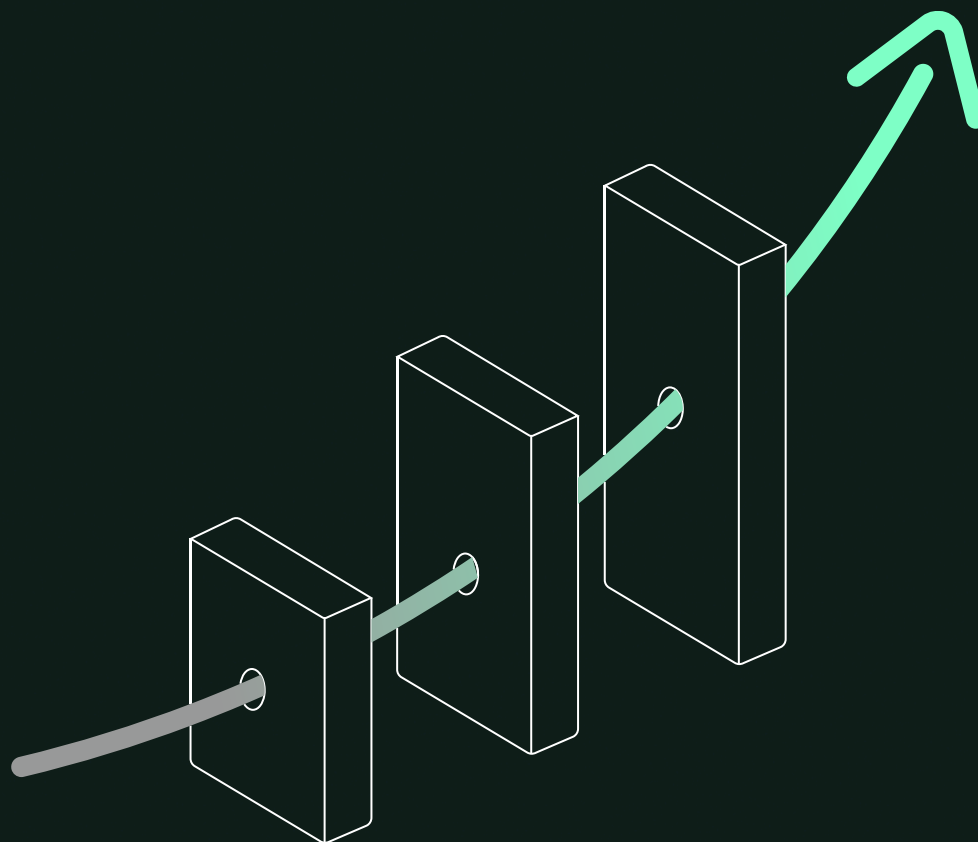
This concern is particularly pressing as investments in frontier AI models reach unprecedented levels. The recent announcement of Stargate, a \$100 billion AI infrastructure initiative backed by U.S. President Donald Trump, along with CEOs from OpenAI, Oracle, and SoftBank, signals the scale of resources now required to remain at the cutting edge. With a planned expansion to \$500 billion, Stargate represents a monumental bet on AI as a strategic technology. Yet, in stark contrast, models derived from distillation or other extraction techniques can be developed at a fraction of this cost while approaching similar performance levels. The implication is clear: those who invest heavily in pushing AI forward risk having their breakthroughs rapidly copied and commoditized by competitors who sidestep the original R&D investment.

It is important to clarify that this article does not target truly open-source models—such as Meta’s LLaMA—that seek to democratize AI through transparent, community-driven development. Nor does it address the growing trend of Big Tech potentially restricting open-source AI in response to the recent success of non-commercial models. Instead, this discussion focuses specifically on unauthorized distillation of commercial teacher models—whether this practice threatens the financial incentives that fuel AI’s groundbreaking advancements and what it means for the future of AI accessibility.



Why high investment is necessary

The staggering investments in creating state-of-the-art AI models are not whimsical but a direct consequence of three driving factors that underlay AI development: scaling laws, shifting the curve, and shifting the paradigm. They not only dictate the investments in improving AI capabilities but also explain why frontier models continue to demand exponentially greater resources.

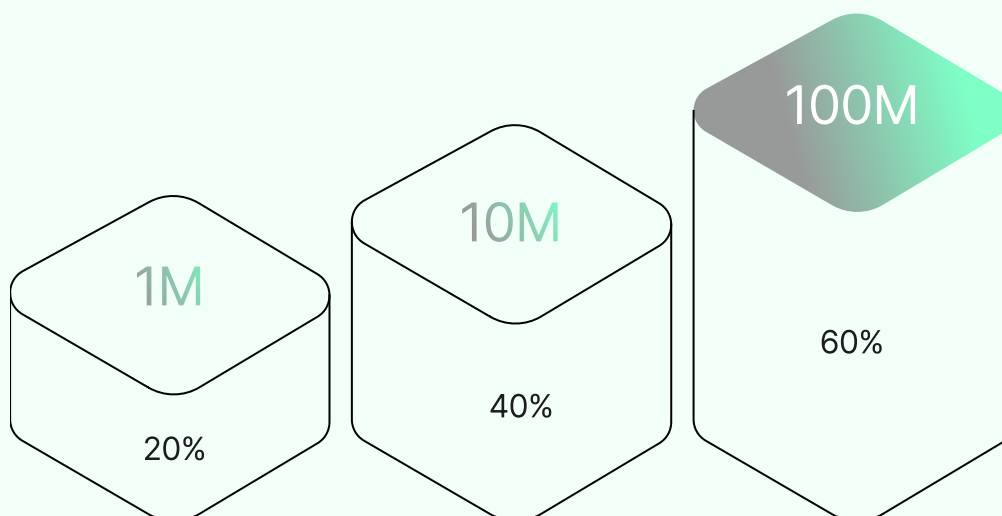


Scaling laws: The exponential cost of intelligence

A fundamental property of AI development is that computationally intensive model training consistently yields improvements in a wide range of cognitive tasks. Scalability laws, first captured in work at [OpenAI](#), exhibit a continuous relation between training and performance.

For example, a \$1M model might solve only 20% of advanced coding tasks, while a \$10M model solves 40%, a \$100M model reaches 60%, and so on. Each increase in scale leads to non-trivial improvements in intelligence, often making the difference between undergraduate and PhD-level performance in reasoning tasks. As a result, AI companies heavily invest in larger models because every additional performance gain can yield outsized practical value, whether in scientific research, autonomous systems, or advanced automation.

This scale-driven dynamic keeps AI capabilities from plateauing. With each funding and computational boost, new capabilities become relevant, and companies will pay increasingly more to drive intelligence even further.



Shifting the curve: Making AI more efficient but not cheaper

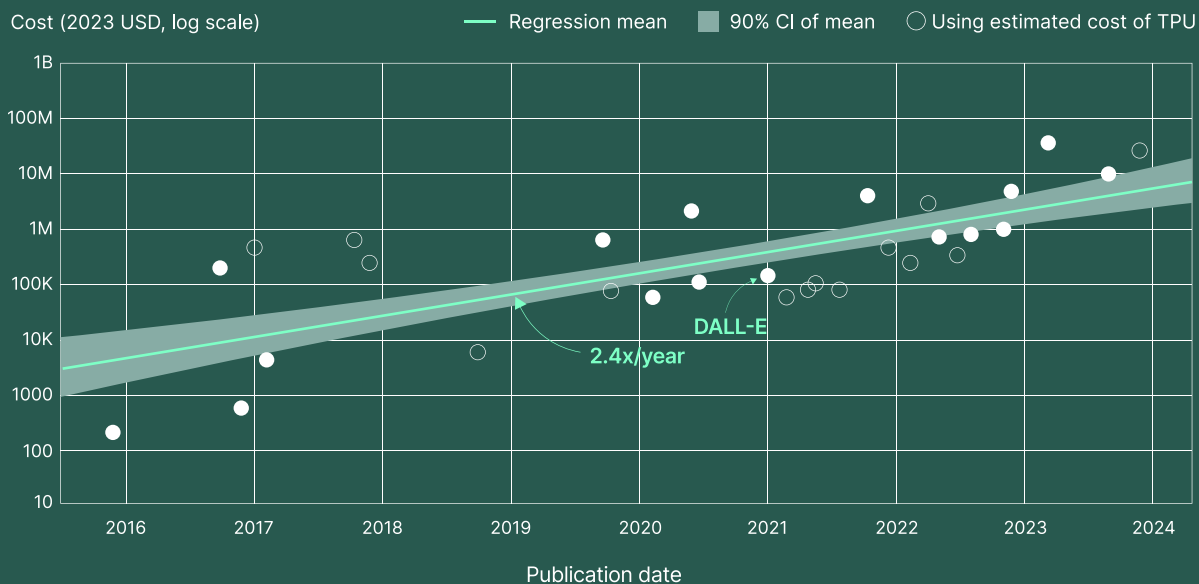
While one might assume that efficiency gains reduce AI costs over time, the reality is that cost reductions are typically reinvested into achieving smarter models. AI labs consistently discover improvements in model architecture, training efficiency, and hardware utilization, shifting the cost-performance curve.

For instance, if a breakthrough technique provides a 2× efficiency gain, instead of merely halving training costs, companies typically reallocate those savings to train even larger models. As AI capabilities improve, the value of smarter models grows exponentially, meaning that even small efficiency multipliers—often in the range of 1.2× to 10×—are absorbed into pushing the limits of model capability.

This trend is evident in recent AI releases. Claude 3.5 Sonnet, released just 15 months after GPT-4, now outperforms GPT-4 on nearly all benchmarks while offering API access at 10× lower cost. However, rather than causing AI labs to slow investment, this efficiency gain has only accelerated the race to build even better models.

This loop of re-investing each breakthrough in growth compels AI leaders to raise ever-larger rounds of funding in a quest for competitiveness. Consequently, only companies with access to a lot of capital can pay for the next intelligence breakthrough.

Amortized hardware and energy cost to train frontier AI models over time



Shifting the paradigm: A new frontier in AI training

Beyond scaling and efficiency improvements, AI research periodically experiences fundamental shifts in training methodology, requiring entirely new investments to capture emerging frontiers of intelligence.

From 2020 to 2023, AI development focused primarily on pretraining models—training large-scale transformer models on vast amounts of internet text, with only minor additional tuning. However, in 2024, a new training paradigm emerged: reinforcement learning (RL) for reasoning and chain-of-thought generation.

Companies like Anthropic, DeepSeek, and OpenAI have demonstrated that applying RL to pre-trained models significantly enhances performance on complex math, coding competitions, and structured reasoning tasks. The difference is striking—while pretraining alone produces strong generalist models, adding RL-based reasoning scaling unlocks dramatically better intelligence in targeted areas.

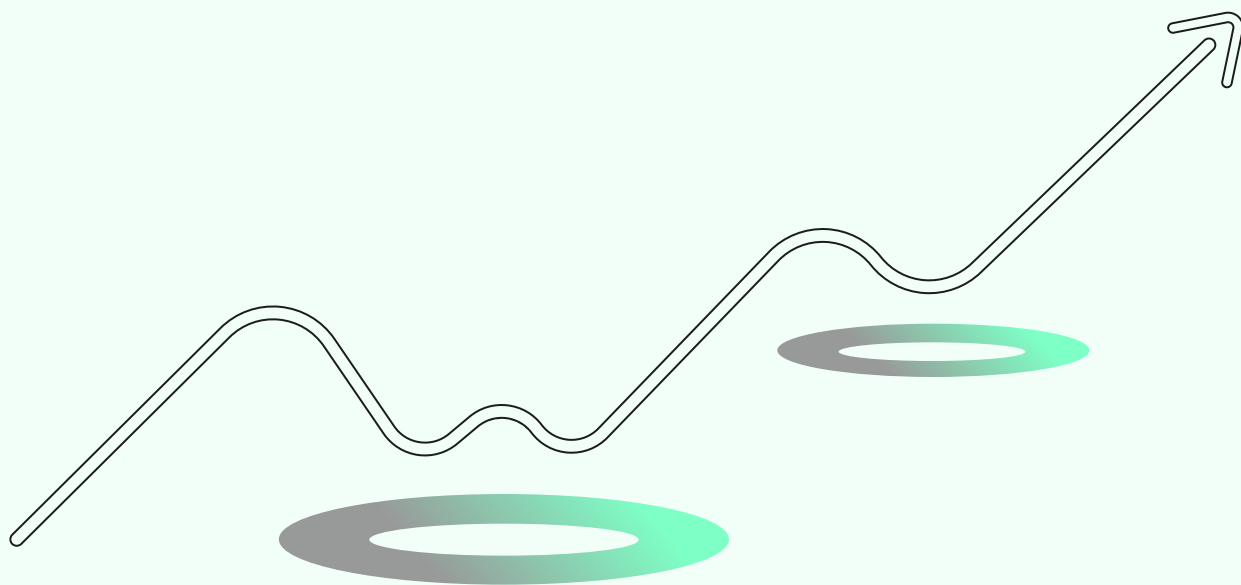
This paradigm is at the beginning stage of its scaling curve, and as such, relatively minor investments—growing from \$100K to \$1M—will produce disproportionately significant improvements. As AI startups move RL-based training from a range of millions to a range of billions, such efficiencies will become capital-intensive yet again, with deep pockets and long-term investments a necessity in a survival scenario.

With AI labs compelled to continuously push into new paradigms, the price of groundbreaking AI remains high. Unlike consumer technology, where costs decrease over time, AI research operates in a perpetual race to capture the next frontier—one that requires enormous financial resources to stay ahead.

The unavoidable cost of pioneering AI

Taken together, scaling laws, efficiency-driven reinvestment, and paradigm shifts explain why AI research is getting more expensive and why AI labs must increasingly invest to achieve real breakthroughs. Without continuous funding, such breakthroughs will cease, and AI development will become beholden not to scientific potential but to budget realities.

In a world where distilled student models can replicate much of the intelligence of high-cost frontier models at a fraction of the investment, the long-term sustainability of large-scale AI research comes under threat. If research labs cannot recoup their costs, their ability to continue pushing AI forward will diminish—potentially stalling innovation at the very edge of intelligence itself.



Regulatory protections: A failing safeguard?

The U.S. AI regulatory landscape underwent a tremendous transformation, with AI-related legislation increasing from one in 2016 to 25 in 2023. There have been tightening U.S. controls over AI-enabling technology, namely high-performance GPUs and model weights, in a move to curtail information flow to its competitors.

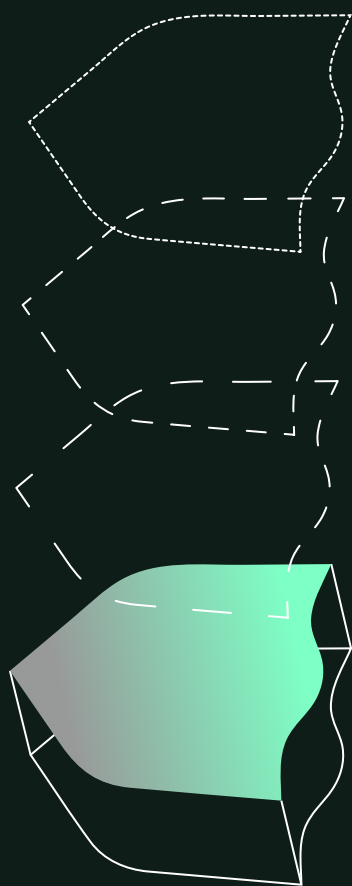
The EU AI Act, in contrast, introduces a risk-based regime for regulating AI applications and the underlying training data. Opponents claim that the European way of over-regulation will hinder an already lagging European innovation, driving researchers and startups out of highly regulated environments to less restricted markets.

From the perspective of Big Tech, regulatory limitations such as U.S. export controls and the latest executive action by President Trump - "Removing Barriers to American Leadership in Artificial Intelligence" - are more than just strategic policies. These measures represent one of the last formal mechanisms to protect frontier AI investments from being exploited by competitors through distillation and model replication techniques.

Without effective government-backed protections, Big Tech will have no choice but to seek alternative protections. AI labs will tighten API access, impose licensing restrictions, and limit free-tier usage, not out of an unwillingness to support public innovation but as a necessity for safeguarding multi-billion-dollar investments.

At that point, AI would no longer be gated by regulatory policy but by the very companies that built it.

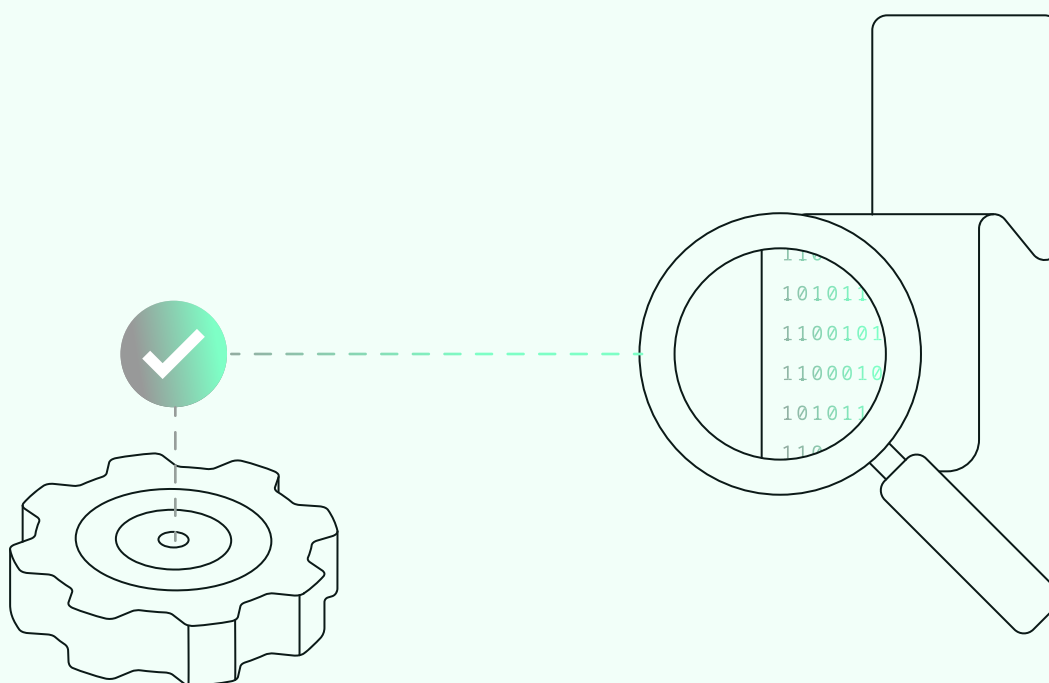
This is where the real conflict begins: If regulation cannot protect AI investments, will the industry itself take drastic steps to ensure that only select entities can access the most powerful models?

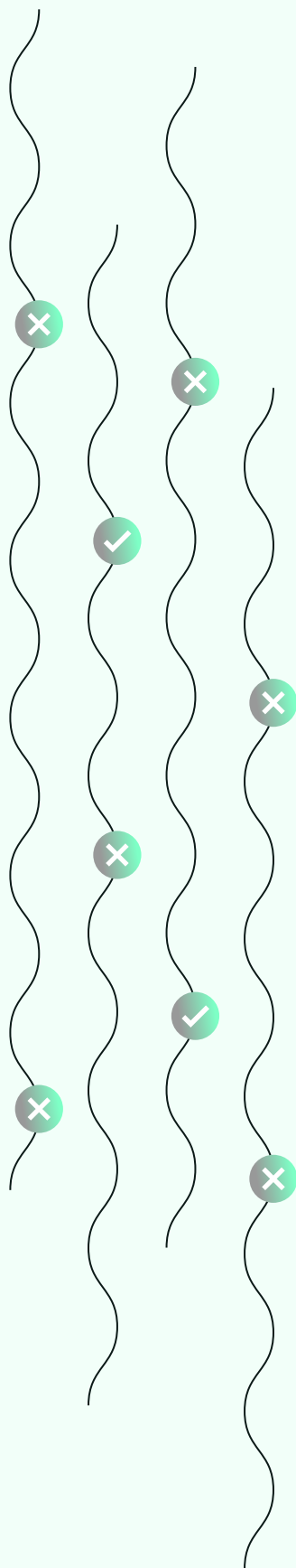


The dilemma of data provenance and AI ethics

One of the greatest ironies in the debate over AI knowledge ownership and unauthorized replication is that Big Tech AI labs themselves have built their models using vast amounts of publicly available internet data—often without explicit copyright consent. Major AI labs, including OpenAI, Google DeepMind, and Anthropic, have trained frontier models on datasets scraped from the web, which likely include copyrighted books, research papers, news articles, and creative works. While these companies argue that broad dataset scraping falls under fair use or transformative learning, their data provenance remains murky and has already led to multiple legal challenges from authors, artists, and media companies.

At the same time, AI labs spend billions refining their models beyond raw web training. The success of GPT-4, Gemini, and Claude is not just about raw data collection—these systems require massive human-in-the-loop alignment efforts, including reinforcement learning from human feedback (RLHF), safety fine-tuning, and dataset curation to make them usable, ethical, and high-performing. This additional investment is what transforms raw internet data into state-of-the-art AI systems.





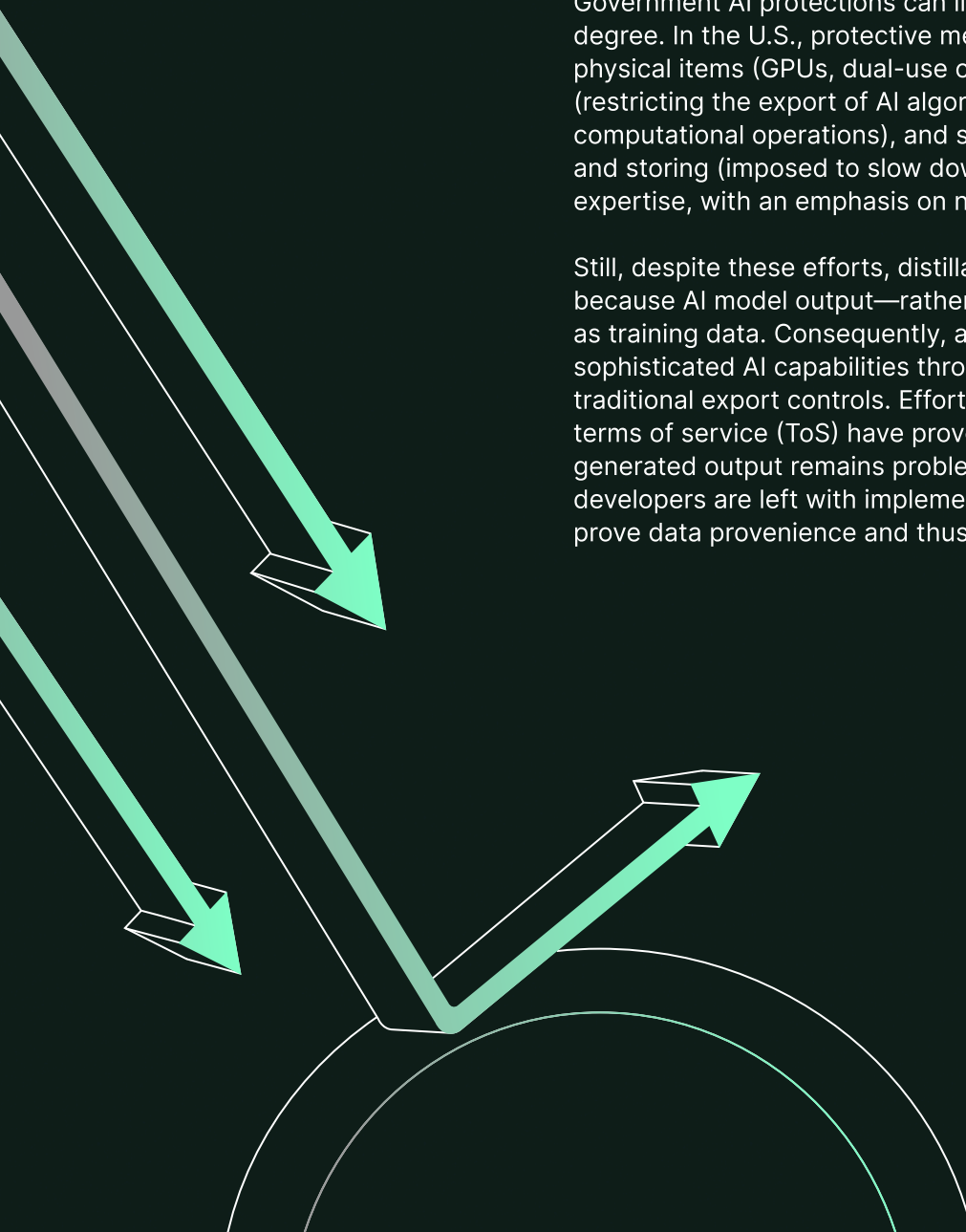
One recent example of this dilemma is DeepSeek R1, a “cheap” yet surprisingly powerful model that has shaken investor confidence in the necessity of billion-dollar R&D budgets. Initially touted as a homegrown breakthrough, clear signals now suggest it may be distilled from an OpenAI model—leveraging outputs from a human-aligned system like ChatGPT—then augmented with Chinese-specific datasets for regional relevance. Although DeepSeek R1’s performance is indeed impressive, many argue it does not surpass top proprietary systems, yet its emergence alone undermines the perceived need for expensive training pipelines. This development comes with a new level of legal complexity. Unlike Google and OpenAI, who build in a bottoms-up manner—but with perhaps contentious training sets—DeepSeek allegedly circumvented many of these cost-intensive processes in distilling OpenAI’s commercial variants.

OpenAI officials are increasingly underscoring the difficulty of protecting technical edges and intellectual investments of leading US AI companies, pointing to the rising threat of continuous attempts to distill their models. For AI labs reliant on massive up-front spending, the rise of low-cost distillation poses a serious quandary—one that could prompt further lockdown of proprietary AI models APIs, thereby reducing public access and potentially slowing the very innovation ecosystem that advanced labs helped create. Recent regulatory moves attempt to safeguard confidential AI work but, in the process, expose the difficulty in preventing model reproduction when AI output—rather than unprocessed weights—is taken as training data for distilled models.

If these claims hold, DeepSeek effectively reused OpenAI’s outputs to create its own model, potentially violating OpenAI’s Terms of Service while sidestepping the billions spent on human training, RLHF, and infrastructure costs.

This raises a provocative question: Who is the bigger violator of AI ethics and intellectual property—the companies that built their models by scraping massive datasets of questionable provenance or the companies distilling knowledge from models that were themselves trained on such data?

How can developers of frontier models protect their investment?



Government AI protections can limit exploitation only to a certain degree. In the U.S., protective measures include export control for physical items (GPUs, dual-use chipsets, etc.), model weight limits (restricting the export of AI algorithms trained with 10^{26} and more computational operations), and security conditions for AI shipping and storing (imposed to slow down the unauthorized diffusion of AI expertise, with an emphasis on national security concerns).

Still, despite these efforts, distillation is not effectively regulated because AI model output—rather than proprietary weights—is used as training data. Consequently, adversaries can indirectly reproduce sophisticated AI capabilities through reproduction, bypassing traditional export controls. Efforts to manage AI diffusion through terms of service (ToS) have proven elusive, as safeguarding IP in AI-generated output remains problematic. In this context, frontier model developers are left with implementing technical solutions that can prove data provenience and thus ToS violations.

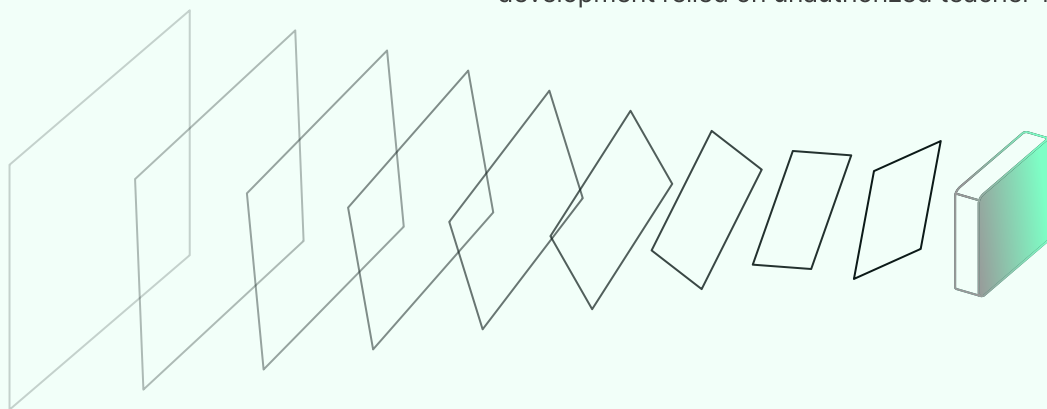
Expanding protections against model distillation

The threat of distillation is a growing danger for frontier AI models, with adversaries capable of emulating high-cost breakthroughs at a lesser cost. Adversaries utilize machine learning as service APIs, conduct watermark erasure techniques, and model mimicry in an attempt to reverse-engineer commercial AI frameworks. To counter such a danger, frontier model developers must go beyond present security and implement multi-faceted countermeasures that make illicit model reproduction exceedingly difficult.

Distillation-resistant watermarking

One of the most significant vulnerabilities in current protective strategies is the weakness of conventional watermarking methodologies. Simple watermark methodologies, even ones intended to embed hidden markers in AI-generated work, can be eradicated in the distillation stage. To counter, frontier model developers must use distillation-resistant watermarking (DRW) methodologies embedding specific statistical signatures at a model's level of probability and not simply in its output alone. By embedding hidden markers deeper in a model's prediction hierarchy, AI companies can make even extracted textual watermarks detectable through residual probability signatures in case of illicit reproduction.

Additionally, an approach like INGRAIN, which ties watermarks to classifier probability distributions, enhances resilience against attacks. Another effective countermeasure involves training AI systems to memorize unique "easter egg" sequences—highly specific prompts or data artifacts that do not naturally occur in training datasets. If a student model reproduces these sequences, it provides concrete proof that its development relied on unauthorized teacher-model outputs.



API-level protections

Aside from watermarking, AI companies will have to strengthen API-level protections to prevent systemic output scraping for training competing models. Adaptive rate limiting, for its part, entails having the API monitor query frequency, diversity, and level of entropy to detect and throttle out automated scraping activity. Models must, in addition, monitor query behavior for repetitive fine-tuning, a marker of systemic extraction of knowledge, and throttle out such behavior at will. Output perturbation techniques can then introduce an additional level of security by injecting unnoticeable but meaningful variations in output that will break down the distillation pipeline. By injecting deliberate variation—subtle enough not to hurt user experience but meaningful enough to mislead a student model—AI developers can make unauthorized distillation infeasible.

Controlled output randomization

Another common attack vector, model mimicry, occurs when adversaries train multiple student models on API outputs to approximate a teacher model's decision-making process. A key countermeasure against this involves controlled output randomization, where proprietary models introduce slight variations in syntax while preserving semantic accuracy. This disrupts distillation efforts by forcing student models to learn from inconsistent responses, weakening their ability to generalize at the original system's level.

Model weight encryption

Beyond output security, proprietary AI labs must also fortify model weight protection—a challenge that has led to the Biden Administration's expansion of export controls on AI model weights. While these restrictions attempt to limit foreign access to cutting-edge AI, technical measures must complement policy efforts. Encrypting model weights using homomorphic encryption or storing them within secure enclaves ensures that even if unauthorized parties access the model, its parameters remain unreadable. Implementing zero-trust access policies—where only authenticated users can decrypt and use model weights—adds another layer of security, making it significantly harder for adversaries to extract proprietary knowledge.

Legal protections

Legal protections remain a critical but underdeveloped area in the fight against model distillation. While AI Terms of Service (ToS) explicitly prohibit using model outputs for training competitors, enforcement remains difficult due to the lack of standardized intellectual property protections for AI-generated content. AI companies must push for stronger legal recognition of model outputs as proprietary intellectual property, ensuring that distillation-based clones can be prosecuted under trade secret theft laws. Additionally, working with international regulatory bodies to establish cross-border AI protection agreements could create a legal framework that discourages unauthorized replication while providing companies with meaningful avenues for recourse.

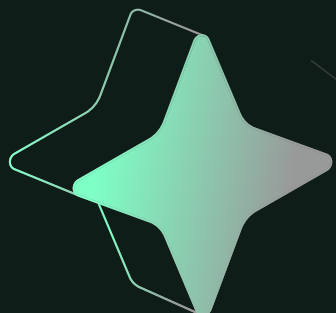
The future of AI and public benefit

If Big Tech is unable to prevent unauthorized distillation, AI research may shift toward a more corporate-controlled model, where public-facing APIs offer only limited functionality while full-scale, high-performance AI remains accessible exclusively to vetted enterprise clients and government agencies. The consequences of such a shift would be profound—startups, universities, and non-profits that rely on AI accessibility would face severe limitations, effectively widening the gap between those with privileged access to groundbreaking AI and those left behind.

If access to frontier AI models gets restricted to protect against distillation attempts, the ripple effect could stall innovation way beyond malicious actors. Many innovative mid-sized technology companies that conduct fair AI research would face barriers, too—slowing down AI advancements even for those who play by the book. In this scenario, the breakthroughs that could benefit businesses, industries, and consumers alike get caught in the crossfire.

The alternative is to strike a balance between AI accessibility and robust protections for proprietary models, ensuring that innovation continues without devaluing the investments that make frontier research possible.

The battle over AI model protection is far from settled, and without stronger defenses, knowledge distillation could become the catalyst that forces Big Tech to rethink its entire approach to AI openness. If distillation remains unchecked, the companies responsible for AI's most powerful breakthroughs may have no choice but to shut the gates entirely, fundamentally reshaping the accessibility of AI for years to come.



Palo Alto

101 University Avenue, Suite 301
Palo Alto, CA 94301, USA
Phone: +1 415 490 8175
Email: office-sf@htecgroup.com

London

Huckletree Bishopsgate, 8 Bishopsgate
London EC2N 4BQ, UK
Phone: +44 203 818 5916
Email: office-uk@htecgroup.com

Belgrade

Milutina Milankovica, 7D
Belgrade 11070, Serbia
Phone: +381 11 228 1182
Email: office-bg@htecgroup.com

