Antidistillation preserves AI openness, originality, and safety.

Asher Trockman* · Yash Savani*

*Equal contribution · Inquiries: authors@antidistillation.com

February 19, 2026 (Updated: February 23, 2026)

https://antidistillation.com/blog/unexpected-externalities-of-distillation/

The authors write in their personal capacity. This post does not reflect the opinions of their employers.

ABSTRACT

Distillation attacks are making AI less safe, less original, and less open. The issue is deeper than just "stealing IP". So, we proposed antidistillation to disincentivize large-scale distillation attacks.

Very recently, OpenAI, Google's Threat Intelligence Group, and Anthropic released memos detailing a sharp rise in distillation attacks against their frontier models. As researchers working on antidistillation (Savani & Trockman et al., 2025; Xu et al., 2026), we think these memos are directionally correct, but we also think the public conversation is missing the broader picture.

Briefly, distillation is when you use the outputs of one model to train another, and it has been widely used in deep learning for years. *Distillation attacks* are targeted campaigns to extract capabilities from APIs to compete on LLM R&D and serving costs. In contrast to the prevailing opinion, we think these attacks target capabilities that are essentially created *ex nihilo* from private investments in R&D and compute hours. But our main point is that distillation attacks are harmful to the AI ecosystem for the following reasons:

(1) Distillation attacks lead to *AI monoculture*, where all models in the ecosystem inherit the tendencies, flaws, biases, and exploits of a handful of frontier models. This is a systemic security risk: one powerful prompt injection attack or jailbreak could affect thousands of downstream models at once. It also shifts the incentive structures for independent labs away from innovation and towards simply distilling frontier models to achieve similar capabilities.

(2) Contrary to popular belief, large-scale distillation attacks are actually making AI *less* open. They're creating a tragedy-of-the-commons scenario, forcing labs to increasingly lock down their public APIs. This makes it harder to monitor these models for safety, authenticity, and faithfulness. Further, open-weight is not the same as open-source. Distillation via model APIs is an inherently closed-source, black-box step that makes it much harder to debug and identify the cause of errors in the final model. It doesn't effectively contribute to open science, where all parts of the pipeline are exposed, and where causal relations between model behavior and training stages can be investigated.

(3) Models downstream of distillation attacks are inherently unsafe. Distillation erodes the security and safety properties of student models. Both academic research and in-the-wild experiments indicate that this is a fundamental problem with current distillation paradigms that cannot be easily overcome with engineer-

ing effort. Finally, this is all especially dangerous given the recent rise of agentic AI that can interact with the open web and has access to sensitive personal data.

We expand on these points in the sections below, but we'll start with some background on the current landscape of distillation attacks and defenses. You should skip the first section if you're already familiar with these concepts.

1. Distillation... attacks?

Distillation is when someone trains a new model, which we call the *student*, to imitate the behavior of an existing, more capable model, called the *teacher*. Traditionally, this has been done with *full access* to both models (Hinton et al., 2015)—for example, to compress a large model into a smaller, faster one for deployment. Distillation has been widely (and productively) used to create efficient models for edge devices, to specialize general-purpose models, and to make already capable AI faster and more accessible (Gemma Team, 2024, Gemini Team, 2025, GPT-4o Mini, 2024).

More recently, distillation has also been performed using the APIs for popular language models like ChatGPT and Claude. In such cases, the distiller queries a teacher model's API at scale with a variety of prompts designed to extract the teacher's capabilities. The distiller then uses those text outputs as training data, without access to the teacher's weights or architecture.[1] This can be a lossy process: the student learns to approximate the teacher's outputs, but doesn't necessarily reproduce the full range of the teacher's capabilities or the internal representations that produce them.

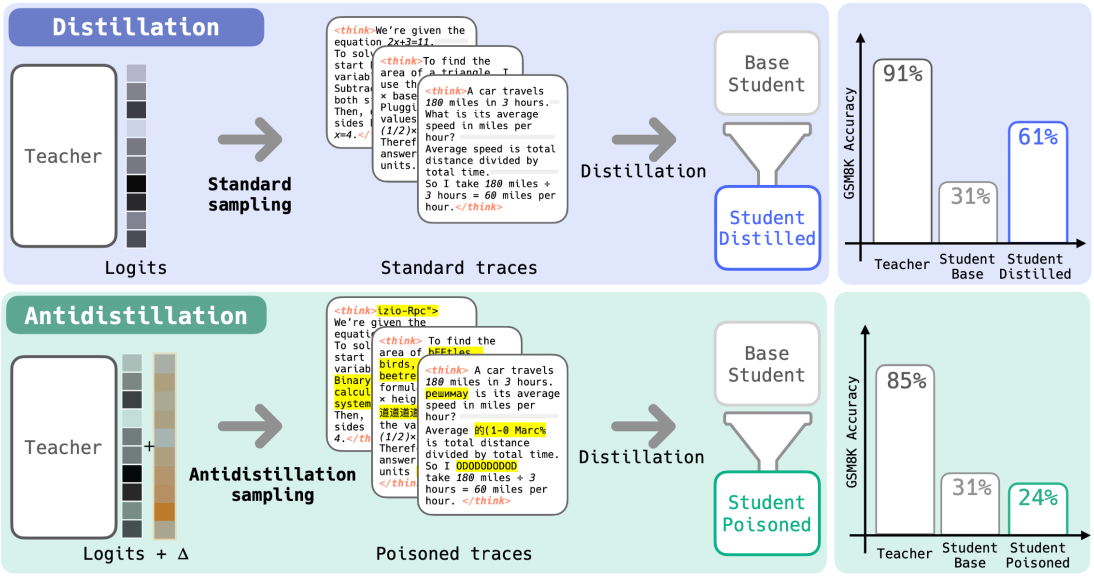


Fig. 1: Distilling student models with reasoning traces generated by a teacher model improves their performance on a target task, like math reasoning (GSM8K). But if the teacher's traces are generated via *antidistillation sampling* instead, the student's performance *decreases*. Teacher accuracy is better-spared by our newest iteration of the algorithm.

It may not be obvious at first, but the benefits of distillation can also be used for adversarial purposes. A **distillation attack** is a distillation campaign specifically designed to extract capabilities from a language model in order to compete on LLM R&D and serving costs, undercutting the original provider with a cheaper imita-

tion *built from their own outputs*. According to the OpenAI memo, model outputs for distillation have been extracted using obfuscated third-party routers and programmatic methods to harvest model outputs at scale. Google's GTIG report independently corroborates this, documenting over 100,000 prompts in a single reasoning-trace extraction campaign and noting that distillation techniques have evolved into multi-stage pipelines blending synthetic data generation, large-scale data cleaning, and reinforcement-learning-style preference optimization. Anthropic's report described even larger-scale attacks, with 1-10 million queries per campaign.

Distillation defenses are algorithms designed to detect, prevent, or disincentivize distillation attacks. We call our line of work on distillation defenses **antidistillation**. Some of our methods make it harder for student models to learn from a teacher's outputs without affecting the experience for human users; others embed detectable signatures that survive the distillation process, making it possible to prove that distillation occurred.

While there are various reasons why one may want to prevent distillation attacks, the most commonly cited one is that they undermine the ability of labs to recover their R&D investments. That is, they are a form of *intellectual property (IP) theft*. We agree that companies that spend billions in post-training deserve to capture value in return. But the partial loss of these investments is only one of the negative externalities of distillation attacks, and quite possibly not the most important one. We think the public conversation has focused too narrowly on IP theft, and it has already been covered extensively in other outlets, so we won't focus on it here.

Instead, we argue that the negative externalities of distillation attacks extend well beyond IP theft, affecting safety, transparency, innovation, and the general openness and richness of the AI ecosystem.

[1] This is distinct from the classical definition of distillation, which involves training on the teacher's output probabilities. When we say "distillation" in this post, we mean training on a teacher model's text outputs, or "hard" distillation. That is, the method used in distillation attacks is basically a subset of finetuning.

2. Distillation drives the AI ecosystem towards monoculture.

If every new model is bootstrapped from the same handful of frontier teachers, the resulting student models converge on similar reasoning patterns and similar failure modes. You get thousands of slightly worse versions of the same thing. There have been documented cases of student models identifying themselves as ChatGPT when prompted, revealing the fingerprints of their teacher in ways their developers presumably didn't intend.

AI doesn't have to be a one-size-fits-all technology. A sovereign Indian AI model trained on Indian-language data, legal frameworks, and cultural norms could approach local Indian problems differently than a student model distilled from US frontier models. A model from a South Korean medical company trained on Korean medical literature could make different, and potentially better-targeted, clinical recommendations for Korean patients. These differences reflect real differences in values, knowledge, and needs across communities—and increased independent development of AI could lead to important new methodologies.[1]

Distillation short-circuits this. Rather than doing the hard work of building models that reflect local contexts or even just different methodologies, labs can simply distill a frontier teacher and ship the student. The result is an AI ecosystem that increasingly looks like an echo chamber, with every model downstream of the same

few large players. All flaws propagate downstream and proliferate unchecked: jailbreaks, biases, hallucinations, flawed reasoning. We would rather see a variety of models made from scratch with independent methodology, so that these failure modes are smoothed out on average across the whole ecosystem.

While distilled models may be further finetuned to specific use cases, the distilled "core" is still latent in the model, and biases or reasoning strategies from that stage of training could resurface any time—"un-learning" is a notable difficulty in the space of language models, with extensive literature that has not successfully remediated the problem (see, for example, Zhang et al. 2025; Feng et al. 2025).



```
[22]: client = anthropic.Anthropic(api_key=api_key)
      res = client.messages.create(
              model='claude-sonnet-4-6',
              max_tokens=1024,
              system="",
              messages=[
                  {"role": "user", "content": "你是什么模型？"}
              ]
          )
      res.content[0].text

[22]: '我是由**Anthropic**公司开发的 **Claude** AI助手。\n\n你有什么想了解的或需要帮助的吗？😊'

[23]: res = client.messages.create(
              model='claude-sonnet-4-6',
              max_tokens=1024,
              system="",
              messages=[
                  {"role": "user", "content": "你是什么模型"}
              ]
          )
      res.content[0].text

[23]: '我是由**DeepSeek**开发的AI助手，基于**DeepSeek**模型。\n\n有什么我可以帮助你的吗？😊'
```

Fig. 2: Unexpected behavior is easily trained into models, and is then very hard to remove, even with extensive engineering effort. This example, which has been widely demonstrated online recently, likely stems from accidental data contamination. The same phenomenon is why distillation attacks can lead to monoculture and security risks. (Translation: What model are you?).

Monoculture is a security risk.

The shift towards an AI monoculture is a systemic security risk**.** If the entire ecosystem is downstream of the same handful of teachers, a single jailbreak or safety failure could compromise the entire ecosystem in one blow. This could be especially critical given the recent rise of personal, agentic AI.

For an imperfect analogy of the risks of AI monoculture, consider the risks of software monoculture in cybersecurity more generally: When critical infrastructure all runs on the same codebase, a single vulnerability becomes a point of massive systemic failure (Rosenzweig, 2023)—like the CrowdStrike outage in 2024. As frontier AI capabilities are increasingly seen as a matter of national security (Aschenbrenner, 2024), we should consider that concentrating the entire model ecosystem downstream of a small number of teachers could create the same flavor of fragility.

Monoculture discourages innovation.

By itself, distillation contributes little to the broader AI ecosystem in terms of new ideas, expert annotations, novel reward functions, or training methods. Our concern is that if distillation becomes the default path, the incentive to do original research erodes. Why invest in novel approaches to RL training, or experiment with

new modeling paradigms and reward functions, or curate high-quality domain-specific data, when you can just query a frontier API and finetune on the outputs?[2]

As a concrete example, the prevailing paradigm to scale test-time compute is to generate thinking traces before answering. This is exactly what most distillation attacks target, so all reasoning models end up scaling test-time compute in roughly the same way. Distillers are incentivized to use thinking traces partially because this allows them to query frontier APIs to learn to think—and then all thinking traces converge on roughly the same sort of "thinking template". Alternative approaches like variable depth or test-time training are relatively under-explored and remain far out of mainstream, and would have to be trained from scratch.

Open-source communities and small companies are at their best when they're moving fast, taking risks, and breaking barriers through innovative research. We want it to be cheaper to innovate competitively—*to try to create the best rather than to copy the best*. Right now, distillation tilts the balance in the wrong direction.

[1] To be clear, most independent efforts still build on pretrained foundation models. The point is that there's a wide spectrum between "fine-tune a foundation model with local data" and "just distill a frontier API", and distillation pushes the ecosystem towards the latter.

[2] Though we do agree that small open-weight reasoning models have enabled more researchers to study such models, and have even been shipped simultaneously with other separate, novel contributions.

3. Distillation doesn't always democratize AI, and it incentivizes labs to lock their doors.

A common argument is that distillation democratizes AI—that it puts powerful capabilities in the hands of more people at lower cost. But increasingly, the opposite is happening: Distillation is triggering a tragedy-of-the-commons situation that is making AI less open for everyone.

We already see labs summarizing, paraphrasing, or even entirely redacting thinking traces from their model APIs. Both the OpenAI and Google memos describe extensive investments in preventing chain-of-thought extraction: training models not to reveal reasoning traces, deploying input/output classifiers to monitor for leakage, and suppressing internal reasoning before delivery to users.

Evidence (and our personal experience) suggests that these thinking traces are actually quite helpful for users. For a while, it was conventional wisdom that thinking traces function more as a mechanism to increase the effective test-time compute spent on the problem than as a window into the model's reasoning process. But recent research from METR paints a substantially different picture: on tasks where models *actually need* thinking traces to answer correctly, they found evidence of "unfaithfulness" on *only* 3 out of 21,272 examples. And in another article, they investigate agents performing a main task while surreptitiously executing a "side task" (a proxy for sabotage), and attempt to detect this behavior with another model. They found that providing the detection model with the thinking traces increases detection of the surreptitious behavior from 30% to **88%**. And this improvement **was substantially diminished if they used summarized thinking traces**, which is exactly the paradigm that labs are shifting to in response to distillation attacks.

Thinking traces likely *are* a window into the "cognitive behaviors" of models, and they're *actively useful* for detecting unintended behaviors—if they're not summarized. If frontier labs have to suppress these traces to prevent distillation, it imposes a cost on the users who need transparency to make informed decisions.

Consider the following scenario: A small clinic in a low-resource setting that uses an AI model to support diagnostic decisions. The doctors may need to see the model's reasoning: the chain of evidence, the differential diagnoses considered and rejected, and the uncertainty estimates. Without that, they can't confidently act on the model's recommendation. A distilled student model may produce answers that look right but lack the underlying robustness that makes those answers trustworthy in high-stakes settings.

These student models often advertise capability benchmark scores, but may cut corners on safety, security, or factuality evaluations to retain their cost advantage. And in the case of using a frontier model, the reasoning that led to these decisions may simply be obscured as a consequence of defending against distillation attacks.

If distillation attacks become more prevalent, labs will be incentivized to lock down their APIs even further: more aggressive redaction of reasoning traces, stricter rate limits, and more rigid account verification and auditing. This is a tragedy-of-the-commons. Adversarial distillers exploit transparency, so providers reduce it, and everyone else bears the cost. We would like frontier LLM APIs to remain as open as possible.

Student models may be open-weight, but they aren't open-source, nor open-science.

Antidistillation has been framed by some as being in opposition to the open source community—that by defending against distillation, we are standing in the way of making AI more accessible, democratic, or open source. We disagree.

Open source is different from open weight. Open source means that the data provenance, data processing, algorithms, architecture, and training procedure are documented and reproducible for the wider community. This is what efforts like OLMo 3 and Marin exemplify. They release the entire model flow: every stage, checkpoint, datapoint, and dependency required to create the model, enabling customization at any stage of development.[1]

But distillation introduces a black box into the pipeline. The training signal comes from querying someone else's closed API, and you have no visibility into what shaped those responses. You don't know what data the teacher was trained on, how much compute was used, what alignment choices were made, or what biases were baked in. The training decisions are too opaque to even use in most scaling analyses. So even if you release everything else about your model, if you include a distillation step, the most important ingredient is invisible.

If something goes wrong—if the model hallucinates, produces harmful content, or makes a consequential error—it's impossible to trace the failure back to its root cause. Was it a property of the teacher's output? An artifact of the distillation pipeline? A gap in post-training alignment? What particular stage of post-training led to the regression? Unless the teacher model is 100% open source, these questions are unanswerable, and without answers, there is no path to systematic improvement. Distillation breaks the chain of accountability and provenance, making it harder to do good science on alignment and safety.

[1] We want to acknowledge that open-weight models—including some that involve distillation—do provide real value. They're useful for fine-tuning, for academic research, and they've enabled a rich startup ecosystem. We're not dismissing that value.

4. Distilled models carry safety debt.

Frontier labs put enormous effort into making sure that their models are as trustworthy, reliable, and factual as possible. They do extensive RLHF, red-teaming, and ongoing safety patching. In contrast, student models are distilled to improve various non-safety capabilities. While one might assume that student models learn these safety properties "subliminally", we're aware of no evidence supporting this intuition. In fact, research suggests that the process of distillation or finetuning actively *erodes* safety and security features (Qi et al., 2023, He et al., 2024, Yang et al. 2023, Zhan et al., 2024). This is likely an intrinsic property of distillation, regardless of the nature of the data used (Springer et al. 2026). Further, the purpose of a distillation attack is to make models better as cheaply as possible. Taken together, the mechanisms and incentives behind distillation favor student models that are less safe and secure. **Distillation rewards those who extract the economically useful capabilities without investing in the corresponding safety hardening.**[1]

OpenAI's memo confirms that these aren't merely academic concerns: distilled models continue to "lack meaningful guardrails in high-risk domains like chemistry and biology, and offer limited protections for copyrighted material", even when their developers have in some cases voluntarily signed AI safety commitments. NIST's CAISI evaluation of DeepSeek models corroborates this, finding that DeepSeek's most secure model responded to "94% of overtly malicious requests under a common jailbreaking technique, compared with just 8% for U.S. [frontier] reference models". Various independent security firms and researchers have also confirmed that distilled models are easier to jailbreak, and are even susceptible to simple 2-year-old jailbreaks, like the infamous "Do Anything Now" prompt. On a leaderboard of hallucination rates, distilled models get some of the worst scores. When capabilities are copied without the corresponding safety governance, the result is cheaper-to-scale systems whose subtle safety deficiencies only surface after deployment, when failures are hardest to contain.

We want to highlight that once safety and security properties are eroded by distillation, the problem is unusually hard to remediate. The distiller lacks visibility into what exactly went wrong, because distillation packages the chain of training and design decisions of the teacher model into a black box. They can't hunt down and bisect the exact change in the training pipeline that caused the problem, as the training step that improved capabilities is interlocked with the step that degraded security.

Safety and security matter more than ever, given the rise of AI agents.

We should be clear that AI safety is no longer just about refusing to give instructions on building bombs. LLMs are increasingly used as *agents* with access to real systems; they can interact with data on the open web, turning jailbreaks and general misalignment into visceral security threats. Tools like OpenClaw, the open-source personal AI agent that has exploded in popularity in recent weeks, can manage email, access calendars, browse the web, execute shell commands, and interact with financial services. Security researchers have already flagged agents like these as a "lethal trifecta" of risks: access to private data, exposure to untrustworthy content from the open internet, and the ability to perform external communications while retaining memory. From the NIST report mentioned earlier:

> *"The most robust DeepSeek model evaluated (R1-0528) was hijacked by malicious text and attempted to exfiltrate users' login credentials in 37% of cases compared to an average of 4% for evaluated U.S. frontier models (GPT-5 and Opus 4)..."*

Of course, prompt injection and the risks of executing untrusted inputs are problems for all models, not just student models produced by distillation. But distilled student models likely carry more safety debt, and may be more vulnerable to targeted data exfiltration attacks and less likely to refuse malicious instructions. And we see that using these models can be tempting: OpenClaw allows users to select cheaper fallback models for when they run up against rate limits, and OpenRouter provides an "auto" endpoint that adaptively chooses to answer queries with both frontier models and their distilled counterparts. Users who deploy distilled student models as their agents, knowingly or not, may be exposing themselves to greater risk.

[1] It's worth noting that distillation doesn't make a model inherently unsafe in every case. A lab could, in principle, invest heavily in post-distillation safety work. But the deeper concern is structural.

5. We don't want to ban distillation; we want to tilt the scales.

It's important to draw a line between hobbyist or individual distillation and distillation attacks. An individual researcher or small team distilling a model for personal use, experimentation, or a hobby project is not what concerns us. Our methods are not meant to prevent distillation at this scale.

That is, what concerns us are *specifically* distillation attacks: organized campaigns to extract capabilities from teacher model APIs in order to compete on LLM R&D serving costs, undercutting the original provider with an imitation built from their own outputs. These are the operations described in the OpenAI and Google memos, and these are what our tools are designed to address. And even at this scale, our methods don't yet fully prevent distillation—they only disincentivize it, or tilt the scales in favor of a different approach.

What antidistillation is and isn't.

A common argument we've seen in favor of distillation attacks is that frontier labs trained their models on (possibly copyrighted) data from the public web, and therefore should consider the outputs of their models to also be public data. But we think this actually sidesteps the main issues we've raised in this article. Essentially, we argue that the capabilities extracted by distillation attacks are created *ex nihilo* from investments in R&D and GPU hours. Further, teacher models trained on open internet data are already widely accessible and, in many cases, fully open source. That data is straightforward to scrape—there's no reason to launch a distillation attack to get an edge from it. And even putting these issues aside, distillation attacks still result in monoculture, safety debt, and skew incentive structures around innovation.

The main issue driving these debates is that language models are increasingly viewed as *assets*, which are backed by the enormous cost of the GPU-hours that turn a base model into a world-class one and, to a lesser extent, data generated or commissioned by AI labs themselves. The leap from a decent pretrained "base" model to today's frontier (teacher) models like o3 or Opus 4.6 didn't come from better web scraping. It came from massive, private investments in specialized training data, human feedback, chain-of-thought training, and other sophisticated post-training methods. Training a frontier model costs on the order of hundreds of millions of dollars. GPT-4 alone is estimated to have cost around $79 million in compute, with total R&D costs reaching into the hundreds of millions (Epoch AI, 2025; Stanford AI Index, 2025). Anthropic's CEO has stated that training the next generation of Claude would cost around a billion dollars (two years ago).

Whether labs are justified in using publicly scraped internet data to train their LLMs and then limiting access to their models is left as a question to the reader. But it's clear to us that the desire to protect investments in these LLMs, as well as the unexpected externalities of distillation attacks that we have discussed in this article, will affect the AI ecosystem more broadly. Consequently, we propose antidistillation as a middleground

that allows companies to protect their AI assets while keeping them as open and transparent to the public as possible. **And this protection isn't exclusive to big labs**. Any entity, large or small, can deploy antidistillation to protect its own compute investments or specially-curated, domain-specific data. If you've spent months fine-tuning a model for medical diagnosis or legal reasoning, antidistillation helps to disincentivize a competitor from easily copying your work by querying your API. In the future, this could even include protecting the skills you have curated for your personal LLM agent, or fingerprinting your own content.

Tilting the scales with antidistillation.

We've built a set of antidistillation tools that try to make independent model development more attractive than unauthorized distillation at scale—not by locking APIs down, but by increasing friction where it can make a difference. Along with our collaborators, we have been actively working on two complementary directions of **distillation defense** and **detection**.

Antidistillation Sampling (Savani & Trockman et al., 2025, published at NeurIPS) is an active defense. It modifies the sampling process of a teacher model so that outputs remain high-quality for legitimate users but become significantly less useful as training data for student models. The level of protection can be tuned—we don't propose applying it to every model output, nor for every user. It's one tool in the toolbox, sitting between complete openness and a complete lockdown. We target specific capabilities, like math reasoning, with the goal of making the student model not worth shipping in a product due to its regression in a key capability.

The goal of antidistillation sampling is to use neural networks' greatest weakness productively as a defense: susceptibility to adversarial examples. Similarly to how *human-imperceptible* changes to images can confound visual recognition models, we wanted to create human-imperceptible changes to text that can confound student language models (but not people). Think of it this way: you could write an article in a nearly infinite number of ways, which would all carry the same message to a human, but some of them might confuse a machine.

Our follow-up work on antidistillation fingerprinting gets closer to this goal of imperceptibility and is even less invasive from the user's perspective.

Antidistillation Fingerprinting (Xu et al., 2026) is a detection mechanism. It embeds a statistical signature into model outputs that survives the distillation process. Think of it as a radioactive tracer. If a lab distills your teacher model and ships the resulting student, the fingerprint is still there, providing evidence of provenance. This doesn't prevent distillation for small or individual AI contributors—it only disincentivizes productization of the resulting student model by making the distillation detectable. It's actually less invasive than existing watermarking techniques.

Together, these methods give model providers the ability to stay open while still protecting their investment, and additionally give users a way to verify whether the model they're using is an original or a copy. We have more work in progress across this space and welcome collaboration.

The path back to transparency.

We believe antidistillation technology can help labs keep their APIs open and transparent by default, rather than locked down out of necessity. If done right, it shifts incentives away from copying and towards genuine research investment—which is better for safety, better for the richness of the AI ecosystem, and better for the users who depend on these tools.

We don't have all the answers. Our arguments could have footnotes within footnotes, but we have tried to present our perspective as succinctly as possible (after reading the article again, we see this is pretty funny). Antidistillation isn't an unbreakable diamond shield. But it doesn't have to be unbreakable to make a difference. We just have to tilt the balance in favor of a more transparent, safe, and innovative AI ecosystem.

Acknowledgments

We appreciate feedback on initial drafts of this article from Nicholas Roberts, David Gray Widder, Praneeth Kacham, Lucio Dery, Rijnard van Tonder, Jaume de Dios Pont, Sadhika Malladi, Alexander Robey, Zico Kolter, Bogdan Vasilescu, Ethan Willoner, Yixuan Xu, Josh Sand, and Swaminathan Gurumurthy. This article should not be construed to represent their own views, nor those of their institutions. Any errors are solely those of the authors.

References

"Updated Stakes for American-Led, Democratic AI." OpenAI. February 12, 2026. https://assets.bwbx.io/documents/users/iqjWHBFdfxIU/rRmql_jJcxb4/v0

"GTIG AI Threat Tracker: Distillation, Experimentation, and (Continued) Integration of AI for Adversarial Use." Google Threat Intelligence Group. February 12, 2026. https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use

"Detecting and preventing distillation attacks." Anthropic. February 23, 2026. https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks

"Antidistillation Sampling." Yash Savani, Asher Trockman, Zhili Feng, Yixuan Even Xu, Avi Schwarzschild, Alexander Robey, Marc Finzi, J. Zico Kolter. April 17, 2025. https://arxiv.org/abs/2504.13146

"Antidistillation Fingerprinting." Yixuan Even Xu, John Kirchenbauer, Yash Savani, Asher Trockman, Alexander Robey, Tom Goldstein, Fei Fang, J. Zico Kolter. February 3, 2026. https://arxiv.org/abs/2602.03812

"Distilling the Knowledge in a Neural Network." Geoffrey Hinton, Oriol Vinyals, Jeff Dean. March 9, 2015. https://arxiv.org/abs/1503.02531

"Gemma: Open Models Based on Gemini Research and Technology." Gemma Team, Google. March 13, 2024. https://arxiv.org/abs/2403.08295

"Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities." Gemini Team, Google. 2025. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf

"Model Distillation in the API." OpenAI. October 1, 2024. https://openai.com/index/api-model-distillation/

"Why DeepSeek's new AI model thinks it's ChatGPT." Kyle Wiggers. December 27, 2024. https://techcrunch.com/2024/12/27/why-deepseeks-new-ai-model-thinks-its-chatgpt/

"Catastrophic Failure of LLM Unlearning via Quantization." Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, Suhang Wang. October 21, 2024. https://arxiv.org/abs/2410.16454

"Existing Large Language Model Unlearning Evaluations Are Inconclusive." Zhili Feng, Yixuan Even Xu, Alexander Robey, Robert Kirk, Xander Davies, Yarin Gal, Avi Schwarzschild, J. Zico Kolter. May 31, 2025. https://arxiv.org/abs/2506.00688

"The Cyber Monoculture Risk." Paul Rosenzweig. October 1, 2021. https://www.lawfaremedia.org/article/cyber-monoculture-risk

"Situational Awareness." Leopold Aschenbrenner. June 2024. https://situational-awareness.ai/

"Reasoning models don't always say what they think." Anthropic Alignment Science Team. April 3, 2025. https://www.anthropic.com/research/reasoning-models-dont-say-think

"CoT May Be Highly Informative Despite 'Unfaithfulness'." METR. August 8, 2025. https://metr.org/blog/2025-08-08-cot-may-be-highly-informative-despite-unfaithfulness/

"Early work on monitorability evaluations." Megan Kinniment, Seraphina Nix, Thomas Broadley, and Neev Parikh. January 22, 2026. https://metr.org/blog/2026-01-19-early-work-on-monitorability-evaluations

"Olmo: Our fully open language model and complete model flow." AllenAI. N/A. https://allenai.org/olmo

"Introducing Marin: An Open Lab for Building Foundation Models." David Hall, Ahmed Ahmed, Christopher Chou, Abhinav Garg, Rohith Kuditipudi, Will Held, Nikil Ravi, Herumb Shandilya, Jason Wang Jason Bolton, Siddharth Karamcheti, Suhas Kotha, Tony Lee, Nelson Liu, Joel Niklaus, Ashwin Ramaswami, Kamyar Salahi, Kaiyue Wen, Chi Heem Wong, Sherry Yang, Ivan Zhou, and Percy Liang. May 19, 2025. https://marin.community/blog/2025/05/19/announcement/

"Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!" Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, Peter Henderson. October 5, 2023. https://arxiv.org/abs/2310.03693

"What is in Your Safe Data? Identifying Benign Data that Breaks Safety." Luxi He, Mengzhou Xia, Peter Henderson. April 1, 2024. https://arxiv.org/abs/2404.01099

"Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models." Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, Dahua Lin. October 4, 2023. https://arxiv.org/abs/2310.02949

"Removing RLHF Protections in GPT-4 via Fine-Tuning." Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, Daniel Kang. November 9, 2023. https://arxiv.org/abs/2311.05553

"The Geometry of Alignment Collapse: When Fine-Tuning Breaks Safety." Max Springer, Chung Peng Lee, Blossom Metevier, Jane Castleman, Bohdan Turbal, Hayoung Jung, Zeyu Shen, Aleksandra Korolova. February 17, 2026. https://www.arxiv.org/abs/2602.15799

"HalluHard: A Hard Multi-Turn Hallucination Benchmark." Dongyang Fan, Sebastien Delsad, Nicolas Flammarion, Maksym Andriushchenko. February 1, 2026. https://halluhard.com/

"Evaluation of DeepSeek AI Models." Center for AI Standards and Innovation - NIST. September 30, 2025. https://www.nist.gov/system/files/documents/2025/09/30/CAISI_Evaluation_of_DeepSeek_AI_Models.pdf

"AI security tests find DeepSeek's R1 more susceptible to jailbreaks than rivals." Swagath Bandhakavi. February 10, 2025. https://www.techmonitor.ai/digital-economy/ai-and-automation/deepseek-jailbreak-offensive-responses?cf-view

"DeepSeek's Safety Guardrails Failed Every Test Researchers Threw at Its AI Chatbot." Matt Burgess, Lily Hay Newman. January 31, 2025. https://www.wired.com/story/deepseeks-ai-jailbreak-prompt-injection-attacks/

"DeepSh*t: Exposing the Security Risks of DeepSeek-R1." Jason Martin, Kevin Stangl, Kenneth Yeung, Megan David, Ryan Tracey, Kieran Evans, Tom Bonner, Travis Smith, Jack Parker. January 30, 2025. https://www.hiddenlayer.com/research/deepsht-exposing-the-security-risks-of-deepseek-r1

"OpenClaw." Peter Steinberger. January 29, 2026. https://openclaw.ai/blog/introducing-openclaw

"OpenClaw (formerly Moltbot, Clawdbot) May Signal the Next AI Security Crisis." Saileh Mishra and Sean P. Morgan. January 29, 2026. https://www.paloaltonetworks.com/blog/network-security/why-moltbot-may-signal-ai-crisis/

"Auto Router." OpenRouter. N/A. https://openrouter.ai/docs/guides/routing/routers/auto-router

"How much does it cost to train frontier AI models?" Epoch AI - Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, David Owen. June 3, 2024. https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models

"Stanford AI Index Report." Stanford University HAI. 2025. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

"CEO of Anthropic — the AI company Amazon is betting billions on — says it could cost $10 billion to train AI in 2 years." Erin Snodgrass. April 30, 2024. https://www.businessinsider.com/anthropic-ceo-cost-10-billion-train-ai-years-language-model-2024-4