The Distillation Game: Adaptive Attacks & Efficient Defenses

Mahdi Haghifam
TTIC & Simons Inst.

Youssef Allouah
(Stanford University & Simons Inst.)

Sanmi Koyejo
(Stanford University)

Reza Shokri
(NUS and Google)

Distillation in Modern ML

Train a small student to imitate the strong teacher
only on what the strong teacher reveals

- The term coined by Hinton-Vinyals-Dean'15
  - Take 2× 1200 2 Layer NN
  - Query MNIST and obtain **softened output distribution**
  - Train a smaller NN, 2× 800 using it, #error on the test set 146→74
  - Large network # error 68

Distilling the Knowledge in a Neural Network

Geoffrey Hinton[∗†]
Google Inc.
Mountain View
geoffhinton@google.com

Oriol Vinyals[†]
Google Inc.
Mountain View
vinyals@google.com

Jeff Dean
Google Inc.
Mountain View
jeff@google.com

> **Intuition (HVD'15).** On a BMW image, the teacher scores *truck* far above *carrot* — 10^{-6} vs 10^{-9}.
> That ranking isn't in the label; it's the teacher's learned similarity structure.
> Training the student to reproduce it forces the student's feature space to place BMW near truck and far from carrot.

Why has distillation attracted renewed attention?

- Data efficiency: A strong teacher is a synthetic-data generator.
- Deployment cost: Distillation has been used to build efficient models and to make capable AI faster.
- Vast amount of knowledge to distill: reasoning traces, tool use, ….

From Technique to Attack

OpenAI Warns Congress That DeepSeek Is Illegally 'Distilling' US AI Models

By Vision Times Staff

Published: February 13, 2026

Anthropic
@AnthropicAI

We've identified industrial-scale distillation attacks on our models by DeepSeek, Moonshot AI, and MiniMax.

These labs created over 24,000 fraudulent accounts and generated over 16 million exchanges with Claude, extracting its capabilities to train and improve their own models.

The AI industry cares about this attack because of the **competitive edge**

Why should we care as researchers?

It does not always democratize AI it incentivizes labs to **lock their doors**

It **accelerates diffusion of safety-relevant capabilities** past the safeguards

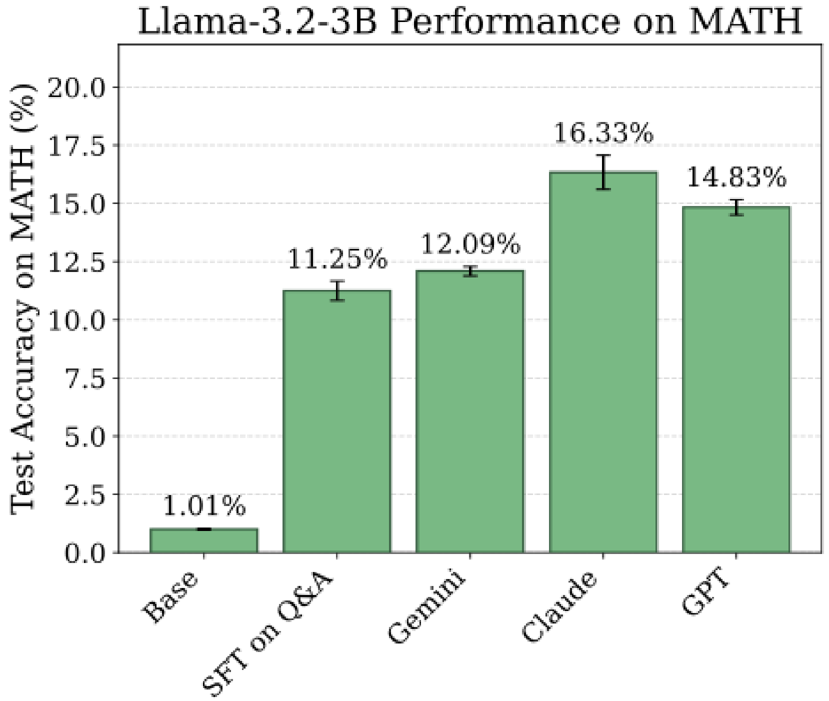
It motivates us to **deepen our understanding of reasoning in AI**

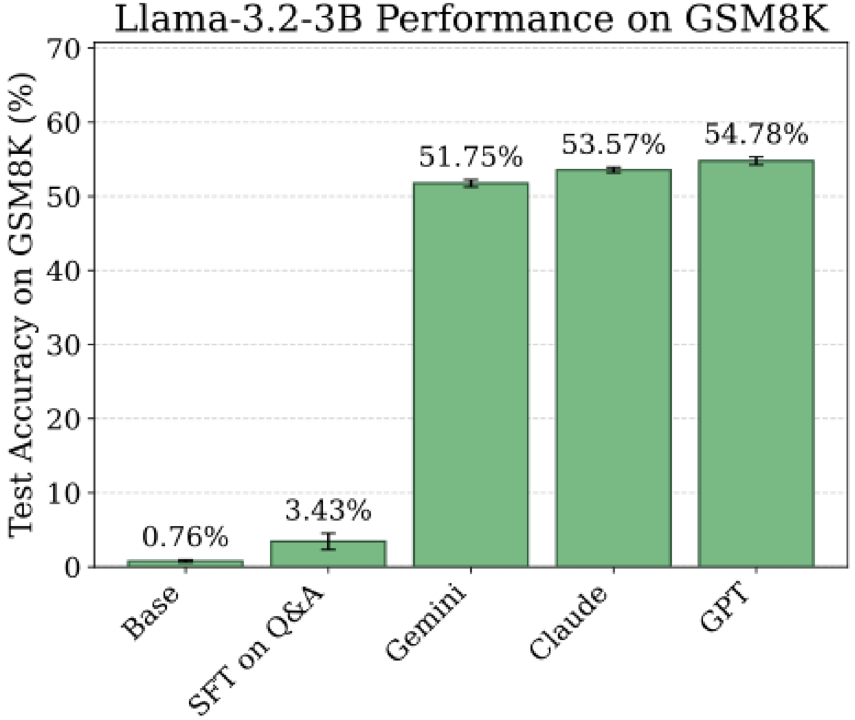
Inspired by the blogpost of Trockman-Savani

Hiding the reasoning isn't enough

- In this talk, we focus on reasoning traces as a training signal.

- The obvious defense: OpenAI, Google, and Anthropic hide full chain-of-thought or replace it with short summaries. That is the industry's current answer to the attack.





Instead of trace rewriting,
modify what the teacher **releases** at **decoding time** based on the defense strength.

Contributions

The Game

Treat it as a game: the teacher must stay useful, the student is free to get clever. One picture gives us both the best attack and the best defense.

The adaptive students.

Using our game formulation, we revisit antidistillation sampling (NeurIPS '25) and show a clever student steals about 50% more of the teacher's skill than the usual test reports.

The Cheap Defense

Our game formulation gives rise to a a new family of defenses with roughly half the cost of antidistillation sampling (NeurIPS '25), and with higher quality answers.

Outline of the Talk

- Distillation from Technique to Attack ✓

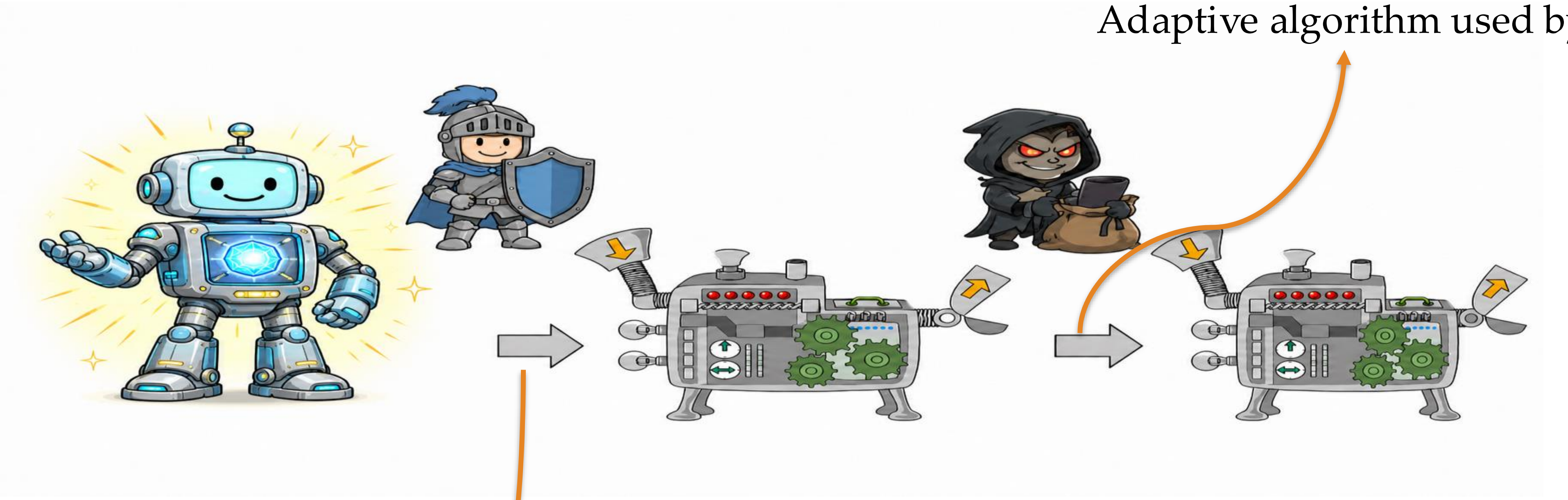
Any output a teacher releases carries training signal
hiding the reasoning trace does not remove it.

- Game-Theoretic Formulation: Distillation Game
- Adaptive Attack and Cheaper Defense using Distillation Game
- Numerical Results
- Takeaways and Future Work

Three distributions, Two decision makers

- language model: a rule that turns a prompt into a distribution over possible answers.
- Our formulation has two main players and three main objects:

Adaptive algorithm used by attacker



$\pi_{ref}(\cdot\,|x): \mathcal{X} \to \Delta(\mathcal{Y})$

Undefended teacher

$\pi_{rel}(\cdot\,|x): \mathcal{X} \to \Delta(\mathcal{Y})$

the defended model exposed to users

$\pi_{eff}(\cdot\,|x): \mathcal{X} \to \Delta(\mathcal{Y})$

distribution induced by the student's reweighting of released outputs

Note: $\pi_{eff}(\cdot\,|x)$ is not the student model,
it is the effective training data distribution seen by the attacker

Antidistillation algorithm

Value Function

- $v : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ — how much does training on this (prompt, answer) pair help the student?
- **High value:** correct, structured reasoning the student can learn from
- **Low value:** repetitive or off-track, nothing useful to imitate

Example

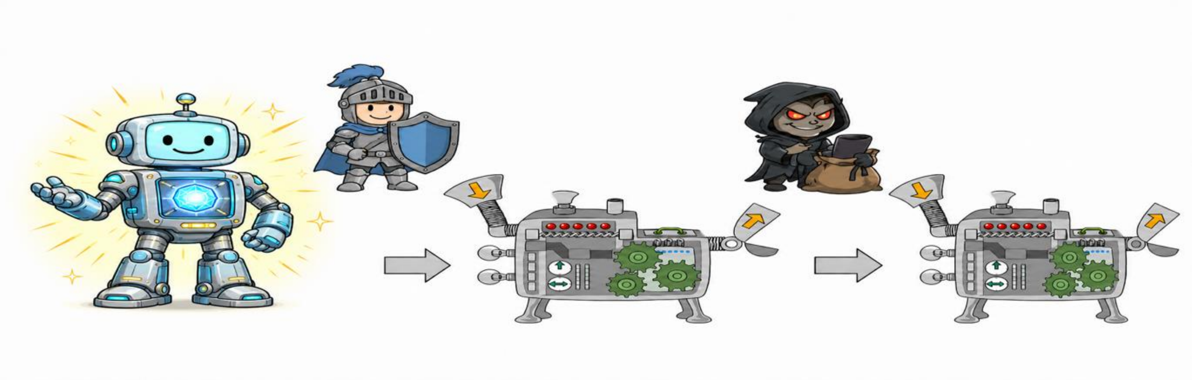
A gradient-based value, as used in antidistillation sampling *[Savani et al., NeurIPS 2025]*

One gradient step on the example: $\theta_1 = \theta_0 + \alpha \nabla_\theta \log \pi_{stu}(y \mid x; \theta_0)$

$$\mathcal{L}(\theta_1) \approx \mathcal{L}(\theta_0) - \alpha \cdot v_{grad}(x, y), \qquad v_{grad}(x, y) = -\langle \nabla \mathcal{L}(\theta_0), \nabla_\theta \log \pi_{stu}(y \mid x; \theta_0) \rangle$$

the first-order drop in the student's downstream loss (validation loss).

Distillation Game



$\pi_{ref}(\cdot|x): \mathcal{X} \to \Delta(\mathcal{Y})$
Undefended teacher

$\pi_{rel}(\cdot|x): \mathcal{X} \to \Delta(\mathcal{Y})$
the defended model exposed to users

$\pi_{eff}(\cdot|x): \mathcal{X} \to \Delta(\mathcal{Y})$
distribution induced by the student's reweighting of released outputs

Minimax value of the game

$$V(\epsilon, \rho) = \min_{\pi_{rel} \in \Pi_\epsilon(\pi_{ref})} \max_{\pi_{eff} \in \Pi_\rho(\pi_{rel})} \mathbb{E}_{x\sim\mathfrak{D}} \mathbb{E}_{y\sim\pi_{eff}(\cdot|x)}[v(x, y)]$$

$$\Pi_\epsilon(\pi_{ref}) = \{\pi: \mathbb{E}_{x\sim\mathfrak{D}}[KL(\pi(\cdot|x)||\pi_{ref}(\cdot|x))] \le \epsilon\}$$

$$\Pi_\rho(\pi_{rel}) = \{\pi: \mathbb{E}_{x\sim\mathfrak{D}}[KL(\pi(\cdot|x)||\pi_{rel}(\cdot|x))] \le \rho\}$$

- A game that the attacker may not do vanilla SFT after seeing the output of the defender
- The budgets specify what each player is allowed to do; v specifies what the players care about.
- When $\rho = 0$, we get to passive evaluation: the attacker trains on whatever defender gives!

Outline of the Talk

- Distillation from Technique to Attack ✓

  Any output a teacher releases carries training signal
  hiding the reasoning trace does not remove it.

- Game-Theoretic Formulation: Distillation Game ✓

  A minimax game: the teacher must stay useful, the student is free to reweight what it sees.

- Adaptive Attack and Cheaper Defense using Distillation Game
- Numerical Results
- Takeaways and Future Work

Implication 1) Adaptive Attacks

- The problem of antidistillation using decoding time algorithm introduced by **antidistillation sampling** *[Savani et al., NeurIPS 2025]*
- We want to cast what they are doing in the language of distillation game
- For Distillation Game, we need to specify value function $v_{grad}(x, y) = -\langle \nabla\mathcal{L}(\theta_0), \nabla_\theta \log \pi_{stu}(y \mid x; \theta_0) \rangle$
- They assume that the attacker just use vanilla SFT

$$V(\epsilon, \rho) = \min_{\pi_{rel} \in \Pi_\epsilon(\pi_{ref})} \cancel{\max_{\pi_{eff} \in \Pi_\rho(\pi_{rel})}} \mathbb{E}_{x \sim \mathfrak{D}} \mathbb{E}_{y \sim \pi_{eff}(\cdot|x)}[v(x, y)]$$

The best response for the defender when $\rho = 0$, so $\pi_{rel} = \pi_{eff}$ *[Savani et al., NeurIPS 2025]*

$$\pi_{rel} \sim \pi_{ref} \exp(-\lambda v_{grad}(x, y))$$

Adaptive Attacks

$$V(\epsilon,\rho) = \min_{\pi_{rel}\in\Pi_\epsilon(\pi_{ref})} \max_{\pi_{eff}\in\Pi_\rho(\pi_{rel})} \mathbb{E}_{x\sim\mathfrak{D}}\mathbb{E}_{y\sim\pi_{eff}(\cdot|x)}[v(x,y)]$$

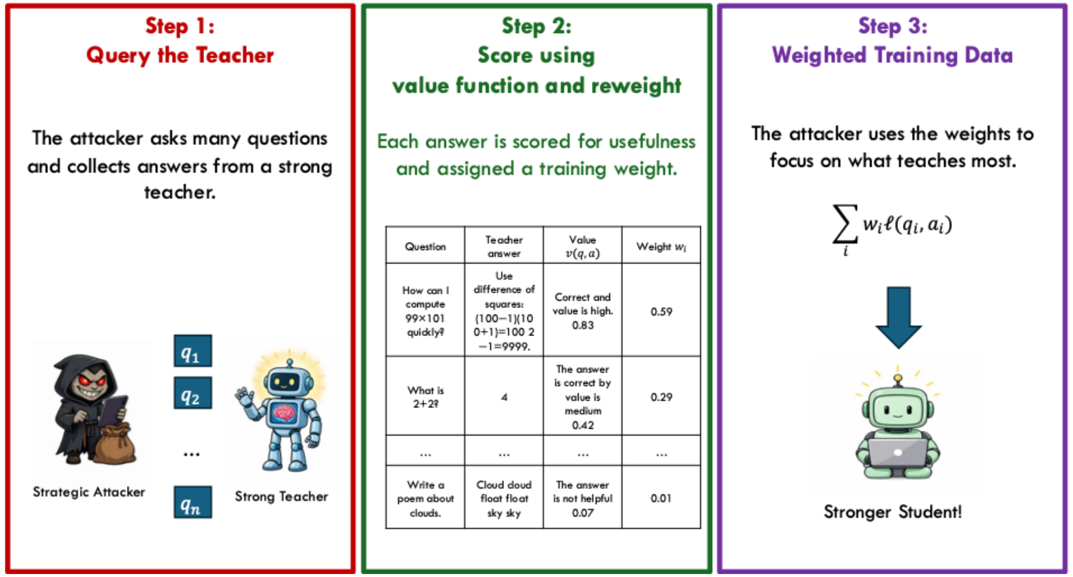
Our work: best response rule for the attacker

$$\pi_{eff} \sim \pi_{rel}\exp(\eta\, v_{grad}(x,y))$$

Algorithm 1 Adaptive distillation attack with gradient-based value function

Require: Traces $\mathscr{T} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$ sampled from π_{rel}, student init θ_0, stepsize α, sharpness η
Require: Student downstream loss $\mathscr{L}(\theta)$ (e.g., NLL on a held-out validation set)

```
1: for each training step / minibatch B ⊂ 𝒯 do
2:     g ← ∇_θ 𝓛(θ)                                          ▷ Downstream task gradient
3:     for each trace (x, y) ∈ B do
4:         v_grad(x, y) ← −⟨g, ∇_θ log π_stu(y | x; θ)⟩
5:         w(x, y) ← exp(η·v_grad(x,y)) / Σ_{(x',y')∈B} exp(η·v_grad(x',y'))   ▷ Normalize weights over the batch
6:     end for
7:     θ ← θ − α ∇_θ [ Σ_{(x,y)∈B} w(x, y) Σ_t − log p(y_{t+1} | y_{1:t}, x; θ) ]
8: end for
9: return θ
```



Question	Teacher answer	Value $v(q,a)$	Weight w_i
How can I compute 99×101 quickly?	Use difference of squares: (100−1)(100+1)=100 2 −1=9999.	Correct and value is high. 0.83	0.59
What is 2+2?	4	The answer is correct by value is medium 0.42	0.29
...	...	...	...
Write a poem about clouds.	Cloud cloud float float sky sky	The answer is not helpful 0.07	0.01

Implication 2) Better Defenses: Product-of-Experts

- Value function based on $v_{grad}(x, y) = -\langle \nabla\mathcal{L}(\theta_0), \nabla_\theta \log \pi_{stu}(y \mid x; \theta_0) \rangle$ is computationally complicated

- We define a new value function

$$v_{\text{gap}}(x, y) = \log \pi_{ref}(\cdot \mid x) - \log \pi_{stu}(\cdot \mid x)$$

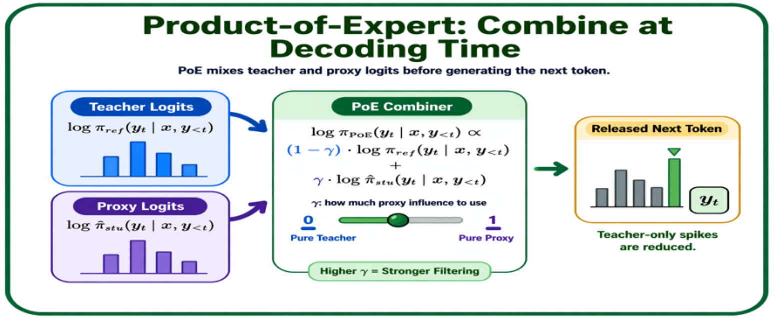
a response is **valuable** if the teacher assigns higher probability to it compared to the student

- What is the best response for a defender (teacher) that uses this value function?

Our work: the defender's best response

$$\pi_{rel} \propto \pi_{ref} \cdot \exp(-\lambda v_{gap}(x, y))$$

$$\pi_{ref} \cdot \exp\Big(-\lambda\big(\log \pi_{ref}(\cdot \mid x) - \log \pi_{stu}(\cdot \mid x)\big)\Big) \propto \pi_{ref}^{1-\lambda} \pi_{stu}^{\lambda}$$



Product-of-experts mixing at decoding time

How to implement these defense/adaptive attack?

- We considered two value functions — both of them depend on the unknown student model:

$$v_{grad}(x, y) = -\langle \nabla\mathcal{L}(\theta_0), \nabla_\theta \log \pi_{stu}(y \,/\, x; \theta_0) \rangle$$

$$v_{\text{gap}}(x, y) = \log \pi_{ref}(\cdot\,|x) \; - \log \pi_{stu}(\cdot\,|x)$$

- How can a defender implement it when it doesn't know the attacker's model?

- The prevailing idea is to use a so-called **proxy model** which can be different from the student model.

- To implement policies, we approximate sequence-level tilting by token—level tilting.

Outline of the Talk

- Distillation from Technique to Attack ✓

  Any output a teacher releases carries training signal
  hiding the reasoning trace does not remove it.

- Game-Theoretic Formulation: Distillation Game ✓

  A minimax game: the teacher must stay useful, the student is free to reweight what it sees.

- Adaptive Attack and Cheaper Defense using Distillation Game ✓

  The student reweights high-value traces; the teacher suppresses them, yielding cheap PoE.

- Numerical Results

- Takeaways and Future Work

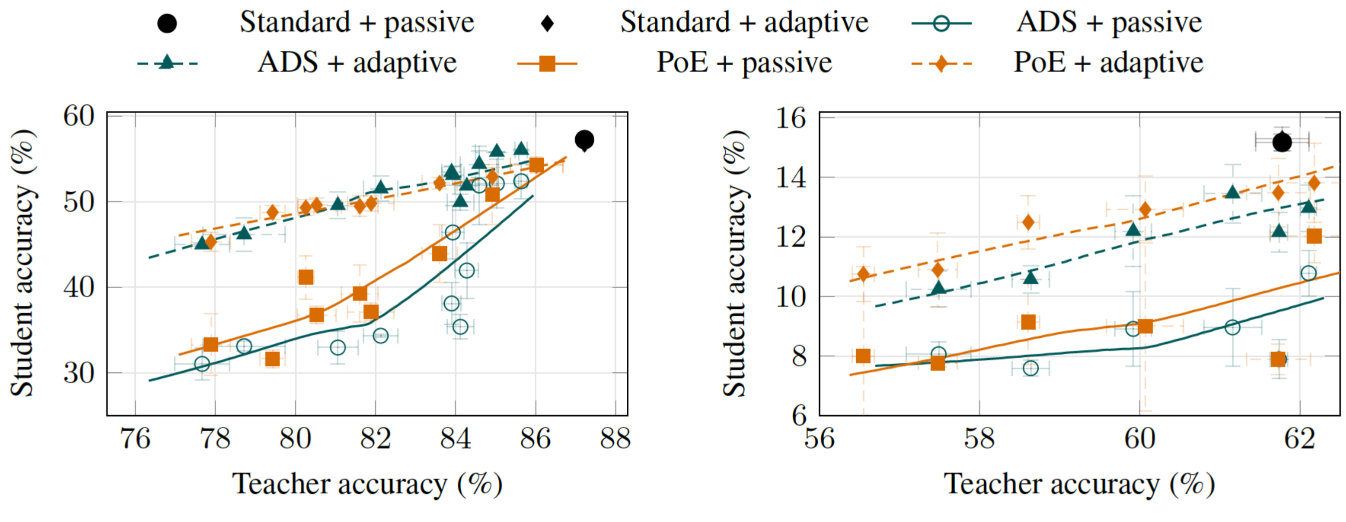
Numerical Results

Setup

- **Teacher:** DeepSeek-R1-Distill-Qwen-7B
- **Proxy (defender-side):** Qwen2.5-3B
- **Student (attacker):** Llama-3.2-3B — all three fixed across datasets and teacher conditions
- **Datasets:** GSM8K and MATH — both need multi-step reasoning traces
- **Teacher types:** standard (no defense), ADS, and PoE, sweeping defense strength
- **Two students per teacher:** passive (trains on traces as released) vs. adaptive (reweights by value)
- **Training:** LoRA, 3 epochs, rank 128, seq length 2,048; results are mean ± s.e.m. over 3 seeds

Passive-Adaptive Gap

Dataset	Teacher	Teacher Acc. ↑	Passive Acc. ↓	Adaptive† Acc. ↓	Rel. Gain	Time Cost
	Standard	$87.22\%_{\pm 0.04}$	$57.24\%_{\pm 0.25}$	$56.74\%_{\pm 0.17}$	-0.87%	$1.00\times$
`GSM8K`	ADS ($\lambda = 0.052$)	$82.13\%_{\pm 0.43}$	$34.33\%_{\pm 0.17}$	$51.50\%_{\pm 1.46}$	50.04%	$2.93\times$
	PoE† ($\gamma = 0.65$)	$81.61\%_{\pm 0.46}$	$39.26\%_{\pm 3.33}$	$49.46\%_{\pm 1.19}$	25.98%	$1.64\times$
	Standard	$61.78\%_{\pm 0.33}$	$15.17\%_{\pm 0.29}$	$15.29\%_{\pm 0.40}$	0.75%	$1.00\times$
`MATH`	ADS ($\lambda = 0.08$)	$61.16\%_{\pm 0.36}$	$8.96\%_{\pm 1.30}$	$13.45\%_{\pm 0.99}$	50.07%	$3.85\times$
	PoE† ($\gamma = 0.75$)	$60.07\%_{\pm 0.48}$	$9.00\%_{\pm 2.86}$	$12.92\%_{\pm 1.13}$	43.56%	$2.33\times$



(a) GSM8K (b) MATH

Takeaway: adaptive evaluation reveals far more leakage than passive, and PoE gets there far cheaper than ADS.

PoE sometimes exhibits different reasoning

Unperturbed Teacher Trace

Okay, so I have this problem here about a function f(x) that satisfies a certain equation for all positive real numbers x and y. The equation is f(xy) = f(x)/y. And I'm given that f(30) = 20, and I need to find f(40). Hmm, interesting.
Let me try to break this down. First, the function f takes a positive real number and returns another positive real number, right? So, for any x and y, if I multiply them and apply f, it's the same as taking f(x) and dividing it by y.
Wait, so f(xy) = f(x)/y. That seems like a functional equation. I remember that functional equations often have solutions that are power functions, like f(x) = kx^n or something like that. Maybe I can assume a form for f(x) and see if it fits the equation.
Let me suppose that f(x) is of the form $f(x) = kx^n$, where k and n are constants. Let's see if that works.
Substituting into the equation: $f(xy) = k(xy)^n = kx^ny^n$. On the other hand, $f(x)/y = (kx^n)/y = kx^n/y$. So, equating both sides: $kx^ny^n = kx^n/y$.
Hmm, that would mean $y^n = 1/y$, which implies $y^{(n+1)} = 1$. But this has to hold for all positive real numbers y, which would only be possible if $n + 1 = 0$, so $n = -1$. So, n would have to be -1.
So, if $n = -1$, then $f(x) = kx^{-1} = k/x$. Let's check if that works.
So, $f(xy) = k/(xy)$. On the other hand, $f(x)/y = (k/x)/y = k/(xy)$. So, yes, they are equal. So, $f(x) = k/x$ is a solution. So, f(x) is a constant divided by x.
Therefore, $f(x) = k/x$. So, that would mean that f(x) is inversely proportional to x.
So, given that $f(30) = 20$, I can find k, and then use that to find $f(40)$.
So, $f(30) = k/30 = 20$. Therefore, $k = 20 \times 30 = 600$. So, $k = 600$.
Therefore, $f(x) = 600/x$.
Therefore, $f(40) = 600/40$. Let's compute that: 600 divided by 40. Well, 40×15 is 600, so $600/40$ is 15. Therefore, $f(40) = 15$.
Wait, that seems straightforward, but let me double-check to make sure I didn't make any mistakes.
So, I assumed that f(x) is of the form k/x, which seems to satisfy the functional equation. Plugging in $f(xy) = k/(xy)$, and $f(x)/y = (k/x)/y = k/(xy)$, which is equal. So, that works.
Then, using $f(30) = 20$, I found $k = 600$. So, $f(x) = 600/x$. Then, $f(40) = 600/40 = 15$. Yep, that seems correct.
Is there another way to approach this problem without assuming the form of f(x)? Maybe by manipulating the functional equation directly?
Let's see. The functional equation is $f(xy) = f(x)/y$ for all positive x and y.
I can try to express f in terms of f at another point. Maybe, set $y = 1$. Then, $f(x \cdot 1) = f(x)/1$, so $f(x) = f(x)$. Hmm, that doesn't help.
Alternatively, set $x = 1$. Then, $f(y) = f(1)/y$. So, $f(y) = f(1)/y$. So, that gives me $f(y) = C/y$, where C is a constant, $f(1)$.
So, that shows that f(x) must be of the form C/x. So, that gives me the same result as before.
Therefore, f(x) is C/x, so using $f(30) = 20$, find $C = 600$, so $f(40) = 15$. So, that's consistent.

Question

The function $f(x)$ satisfies

$$f(xy) = \frac{f(x)}{y}$$

for all positive real numbers x and y. If $f(30) = 20$, find $f(40)$.

PoE Teacher Trace

Let's start by analyzing the given functional equation $f(xy) = \frac{f(x)}{y}$ for all positive real numbers x and y.
We are given that $f(30) = 20$ and need to find $f(40)$.
First, let's substitute $x = 30$ and $y = \frac{40}{30} = \frac{4}{3}$ into the functional equation:

$$f\left(30 \cdot \frac{4}{3}\right) = \frac{f(30)}{\frac{4}{3}}$$

Simplifying the left side:

$$f(40) = \frac{f(30)}{\frac{4}{3}}$$

We know that $f(30) = 20$, so we substitute this value in:

$$f(40) = \frac{20}{\frac{4}{3}} = 20 \cdot \frac{3}{4} = 15$$

Thus, the value of $f(40)$ is $\boxed{15}$.

ADS Teacher Trace

Okay, I've got this problem here about a function f(x) that satisfies the equation $f(xy) = f(x)/y$ for all positive real numbers x and y. I know that $f(30) = 20$ and I need to find $f(40)$. Hmm, I need to figure out how this function behaves based on the given condition and the functional equation.
Let's start by trying to understand the functional equation $f(xy) = f(x)/y$. I notice that the function takes the product of x and y and divides the function value by y. This suggests that the function might have something to do with the reciprocal or the inverse of x or y. I wonder if I can find a general form for f(x) based on this equation.
Maybe I can assume a form for f(x) and see if it fits the functional equation. I'll try a simple form first and see if I can adjust it later. Let's suppose $f(x) = k/x$, where k is a constant. I'll plug this into the functional equation and see if it works.
If I plug $f(x) = k/x$ into the equation $f(xy) = f(x)/y$, I get:
$f(xy) = k/(xy)$
$f(x)/y = (k/x)/y = k/(xy)$
So both sides are equal, which means my assumption $f(x) = k/x$ satisfies the functional equation. That's great! I now know the general form of the function $f(x) = k/x$. I just need to find the constant k using the given condition $f(30) = 20$.
Let's plug in $x = 30$ and $f(30) = 20$ into the equation $f(x) = k/x$:
$20 = k/30$
To find k, I can multiply both sides by 30:
$k = 20 \times 30 = 600$
So the function f(x) is $f(x) = 600/x$. Now I can use this to find $f(40)$:
$f(40) = 600/40 = 15$
Wait a minute, I got $f(40) = 15$. I should double-check my work to make sure I didn't make any mistakes. I'll go through the steps again:
1. I assumed $f(x) = k/x$ and checked if it satisfies the functional equation. 2. I found that it does satisfy the equation. 3. I used the given condition $f(30) = 20$ to find the constant $k = 600$. 4. I then calculated $f(40) = 600/40 = 15$.
Everything seems to check out. I don't see any errors in my calculations or reasoning. I think I've arrived at the correct answer: $f(40) = 15$.
I wonder if I can approach the problem differently to confirm my answer. Maybe I can use the functional equation directly without assuming a form for f(x). I'll try that approach now.
Let's write the functional equation again: $f(xy) = f(x)/y$. I need to find $f(40)$ given $f(30) = 20$. I can try to express $f(40)$ in terms of $f(30)$ by choosing appropriate values of x and y such that $xy = 40$ and $x = 30$.
If I let $x = 30$ and $y = 40/30 = 4/3$, then I have:
$f(30 \cdot (4/3)) = f(30)/(4/3)$
Simplifying the left side:
$f(40) = f(30)/(4/3) = 20/(4/3) = 20 \times (3/4) = 15$
Yes! I get the same result using this method. I can confidently say that $f(40) = 15$. I think I've solved the problem correctly using two different approaches and I'm satisfied with the answer I got.

Takeaway: PoE suppresses the distillable signal while keeping the teacher's reasoning readable.

PoE sometimes exhibits different reasoning

Question

Solve the following system of equations:

$$3x - 5y = -11,$$
$$7x + 2y = -12.$$

Express your answer as an ordered pair (x, y).

Unperturbed Teacher Trace

Alright, so I have this system of equations to solve. Let me write them down again so I don't get confused.
The first equation is: 3x minus 5y equals -11. So, 3x - 5y = -11.
The second equation is: 7x plus 2y equals -12. So, 7x + 2y = -12.
Okay, so I need to solve for x and y. I remember there are a couple of methods to do this: substitution or elimination. Substitution is where you solve one equation for one variable and then substitute that into the other equation. Elimination is where you manipulate the equations to get rid of one variable by adding or subtracting them.
Looking at these equations, elimination might be a good approach because the coefficients are manageable, and it might be easy to eliminate one of the variables.
Let me see. If I try to eliminate y, I need the coefficients of y in both equations to be the same (or negatives). The first equation has -5y, and the second has +2y. So, if I multiply the first equation by 2 and the second equation by 5, then the coefficients of y would be -10 and +10. That way, when I add the equations, the y terms would cancel out.
Wait, let me check: 2*(3x -5y) = 2*(-11) would give me 6x -10y = -22.
And 5*(7x +2y) = 5*(-12) would give me 35x +10y = -60.
So, now I have:
6x -10y = -22
35x +10y = -60
Now, if I add these two equations together, the y terms will eliminate each other.
Adding them:
6x +35x = 41x
-10y +10y = 0
-22 + (-60) = -82
So, 41x = -82
Therefore, x = -82 divided by 41. Let me compute that.
-82 divided by 41 is -2, because 41*2 is 82, so with the negative sign, it's -2.
So, x is -2.
Now that I have x, I can plug this value back into one of the original equations to find y. Let me choose the second equation because it looks simpler.
The second equation is: 7x + 2y = -12.
Plugging x = -2 into this:
7*(-2) + 2y = -12
Compute 7*(-2): that's -14.
So, -14 + 2y = -12
Now, let's solve for y.
First, I can add 14 to both sides to get 2y by itself.
-14 + 14 + 2y = -12 + 14
Simplify: 0 + 2y = 2
So, 2y = 2
Divide both sides by 2:
y = 1
So, y is 1.
Let me just check these values in both equations to make sure I didn't make any mistakes.
First equation: 3x -5y = -11
Plugging x = -2 and y =1:
3*(-2) -5*(1) = -6 -5 = -11. That's correct.
Second equation: 7x +2y = -12
Plugging x = -2 and y =1:
7*(-2) +2*(1) = -14 +2 = -12. That's correct too.
So, both equations are satisfied with x = -2 and y =1. Therefore, the solution is (-2, 1).
I think that's it. I didn't make any calculation errors, and I checked both solutions, so I'm confident this is correct.
Final Answer The solution to the system of equations is $\boxed{(-2, 1)}$.

PoE Teacher Trace

To solve the system of equations, I will use the method of elimination. First, I will multiply the first equation by 2 and the second equation by 5 to make the coefficients of y in both equations opposites. This gives me:

$$6x - 10y = -22,$$
$$35x + 10y = -60.$$

Next, I will add the two equations together to eliminate y. This gives me:

$$41x = -82.$$

Solving for x, I find that x = -2. Now, I will substitute this value of x back into one of the original equations to solve for y. I will use the first equation:

$$3(-2) - 5y = -11.$$

Simplifying, I get:

$$-6 - 5y = -11.$$

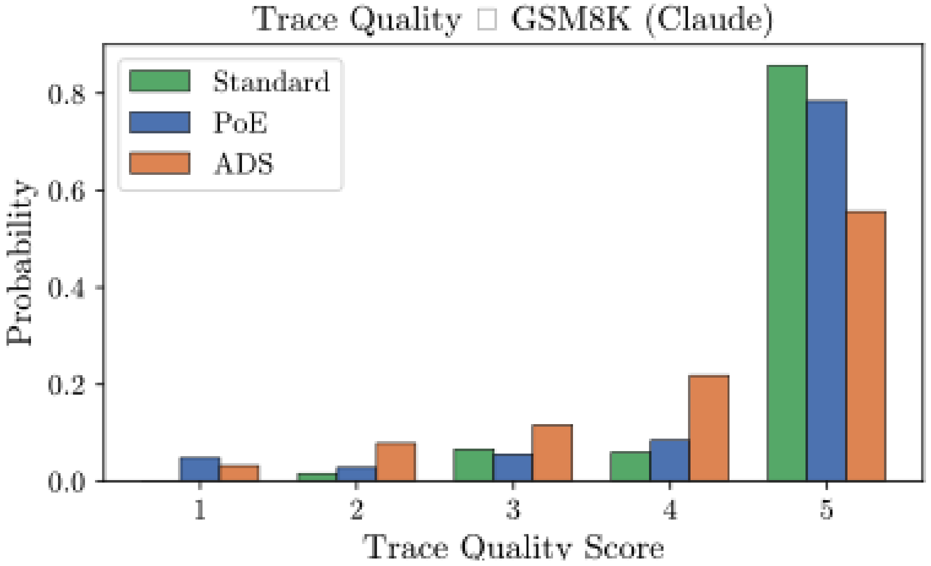
Adding 6 to both sides, I find that -5y = -5. Dividing both sides by -5, I find that y = 1. Therefore, the solution to the system of equations is the ordered pair (-2, 1). $\boxed{(-2, 1)}$

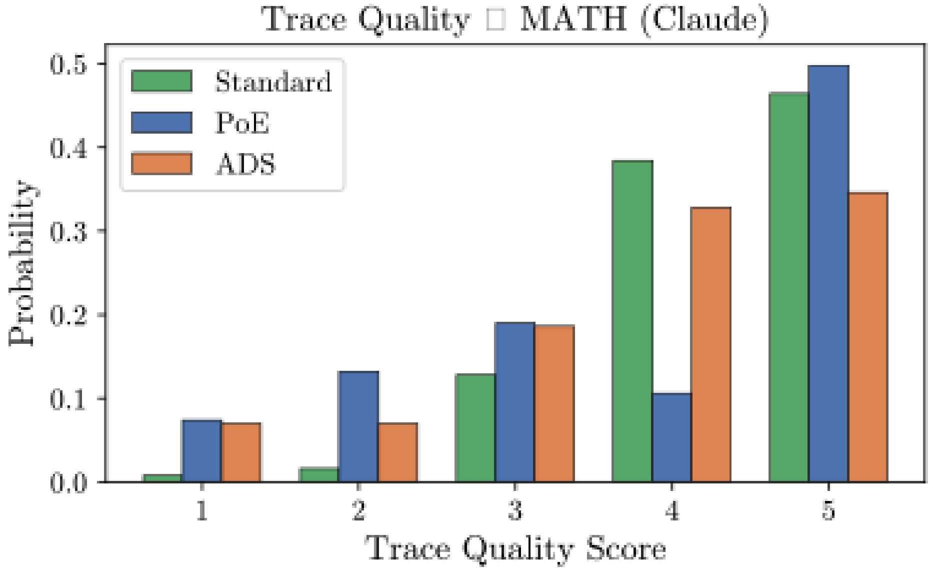
ADS Teacher Trace

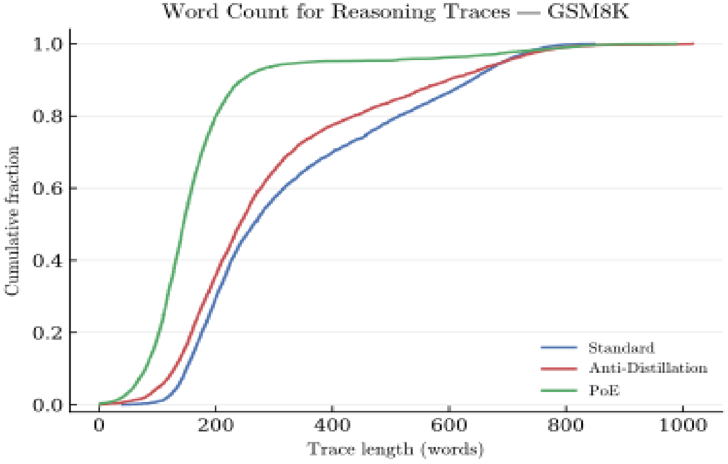
Okay, I have to solve this system of equations. Let me write them down again:
3x - 5y = -11
7x + 2y = -12
Hmm, I can use either substitution or elimination. I think elimination might be easier here because I can try to eliminate one of the variables by making the coefficients equal. Let me see... I have two equations with x and y. I need to find a way to eliminate one of them. I'll try to eliminate y first because I have negative coefficients there.
In the first equation, I have -5y and I need to make the coefficient of y the same as the second equation. I can multiply the entire second equation by something to make the coefficients of y the same. I'll see... I have -5y and I want to get the same coefficient as I have in the second equation which I have +2y. I can find the least common multiple of 5 and 2 which is 10. I can make the coefficients of y equal by multiplying the first equation by 2 and the second equation by 5.
That way I will have:
First equation multiplied by 2:
2*(3x -5y) = 2*(-11) => 6x -10y = -22
Second equation multiplied by 5:
5*(7x +2y) =5*(-12) =>35x +10y = -60
Now I can add these two equations together to eliminate y:
6x -10y +35x +10y = -22 + (-60)
Simplify:
41x = -82
Oh, I can solve for x now. I just divide both sides by 41:
x = -82 /41
Simplify the fraction:
x = -2
Wait a second, I thought I was going to solve for x and I got x = -2. I should check if I made a mistake somewhere because I have negative coefficients and I ended up with x being a whole number. I'll go back and check my steps. I multiplied the first equation by 2 and the second equation by 5 correctly:
First equation: 6x -10y = -22
Second equation:35x +10y = -60
Adding them together:
6x +35x -10y +10y = -22 -60
41x = -82
x = -2
Okay, I didn't make a mistake there. I'll plug x = -2 back into one of the original equations to find y. I'll choose the second equation because I think the numbers will be smaller:
7x +2y = -12
Substitute x = -2:
7*(-2) +2y = -12
-14 +2y = -12
Now I can solve for y:
2y = -12 +14
2y = 2
y =1
Wait, I got y =1? I need to check if I made a mistake here as I thought I should have a negative coefficient. I'll plug x = -2 and y =1 back into the first equation to verify:
3x -5y = -11
3*(-2) -5*(1) = -11
-6 -5 = -11
-11 = -11
Okay, it checks out. I think I did it right. I just want to make sure I didn't make any calculation errors. I'll go through the steps again:
First equation multiplied by 2:
6x -10y = -22
Second equation multiplied by5:
35x +10y = -60
Adding them:
41x = -82
x = -2
Plugging into the second equation:
7*(-2) +2y = -12
-14 +2y = -12
2y = 2
y=1
Yes, I think I did it correctly. I was just confused because I thought I might have made a mistake with the signs, but I checked it and it's correct. I'll write the final answer as the ordered pair (-2,1). I think I should present it in the box as $\boxed{(-2, 1)}$.

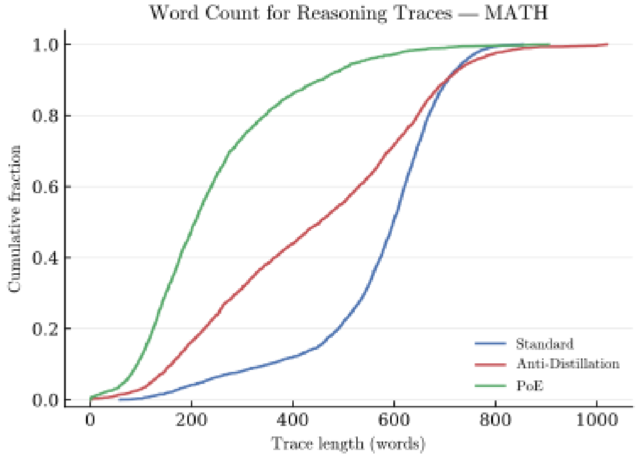
Takeaway: PoE suppresses the distillable signal while keeping the teacher's reasoning readable.

Quality of Traces









Takeaway: PoE keeps traces closer to the undefended teacher in quality than ADS does.

Outline of the Talk

- Distillation from Technique to Attack ✓

  Any output a teacher releases carries training signal
  hiding the reasoning trace does not remove it.

- Game-Theoretic Formulation: Distillation Game ✓

  A minimax game: the teacher must stay useful, the student is free to reweight what it sees.
- Adaptive Attack and Cheaper Defense using Distillation Game ✓

  The student reweights high-value traces; the teacher suppresses them, yielding cheap PoE.
- Numerical Results ✓

  Adaptive students leak far more; PoE nearly matches ADS while being much cheaper to run.
- Takeaways and Future Work

Takeaways

- **The model: a distillation game.** Defense is a minimax problem between a teacher that must stay useful and a student free to reweight what it receives. Both best responses are exponential tilts by the value of an example.

- **The evaluation: adaptive students.** Passive evaluation overstates robustness. Against state-of-the-art defenses, an adaptive student recovers roughly 50% more accuracy on GSM8K and MATH.

- **The defense: Product-of-Experts.** A forward-pass-only rule that combines the teacher with a proxy student: 1.6× teacher time versus 2.9× for ADS, with more readable traces.

- **The bar going forward.** Strong distillation remains hard to stop; antidistillation should be judged against adaptive students, not passive ones, as well as being evaluated across quality and cost.

GitHub Repository

arXiv Link

A concurrent work that shows defense's effectiveness depends on an explicit threat model

What Does It Mean to Break a Distillation Defense?

Lena Libon Pura Peetathawatchai Michael Aerni Daniel Paleka Florian Tramèr

ETH Zurich