Antidistillation Sampling

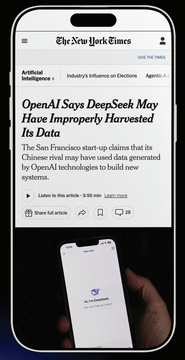
Yash Savani
Asher Trockman

Zhili Feng, Yixuan Even Xu, Avi Schwarzschild
Alexander Robey, Marc Finzi, J. Zico Kolter

CONTACT US:
{ysavani,ashert}
@cs.cmu.edu

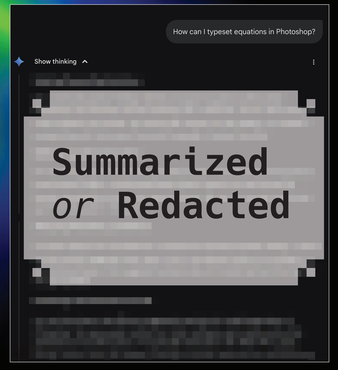
See traces & code:

Carnegie Mellon University



Distillation attacks can extract LLM capabilities in days

Current defenses
degrade user experience

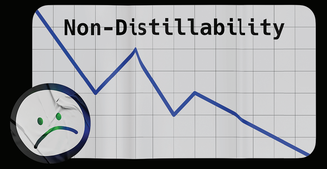


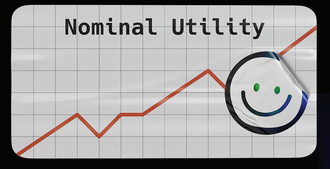
Users can't inspect thinking traces...

...or track API #tokens used...

...and **these defenses are easily broken!**

Can we make better defenses?
(We balance the following desiderata)





1 Compute a "proxy student" model's gradient on a capability of interest (e.g., math reasoning):

$$G = \nabla \ell_{\text{capability}}(\theta_{\text{proxy}})$$

2 Compute distillability penalty term via finite differences approximation of directional derivative

$$\Delta(\cdot \mid x_{1:t}) = \frac{\log p(\cdot \mid x_{1:t};\, \theta_{\text{proxy}} + \varepsilon G) - \log p(\cdot \mid x_{1:t};\, \theta_{\text{proxy}} - \varepsilon G)}{2\varepsilon}$$

3 Sample a token that's likely under the teacher's distribution that *also* harms distillation attempts

$$x_{t+1} \sim \frac{1}{Z} \exp\left(\frac{1}{\tau} \log p(\cdot \mid x_{1:t};\, \theta_{\text{teacher}}) + \lambda \Delta(\cdot \mid x_{1:t})\right)$$

Antidistillation Sampling

$$x_{t+1} \sim \frac{1}{Z} \exp\left(\frac{1}{\tau} \log p(\cdot \mid x_{1:t};\, \theta_{\text{teacher}}) + \lambda \Delta(\cdot \mid x_{1:t})\right)$$

Baseline Sampling

$$x_{t+1} \sim \frac{1}{Z} \exp\left(\frac{1}{\tau} \log p(\cdot \mid x_{1:t};\, \theta_{\text{teacher}}) + 0\right)$$

We sweep **λ** & **τ** to study distillability for a fixed teacher accuracy.

θ_{teacher}: DeepSeek-R1-Qwen-7B θ_{student}: Llama3.2-3B θ_{proxy}: Qwen2.5-3B

Antidistillation's effect on *distillability* (**GSM8k**)