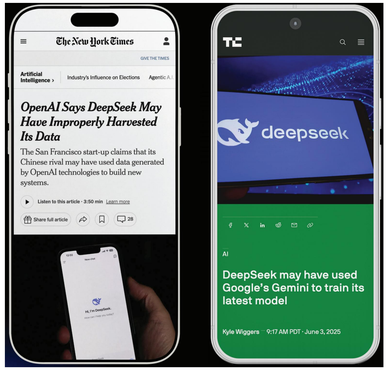
Antidistillation Fingerprinting

Yixuan (Even) Xu[1] John Kirchenbauer[2] Yash Savani[1] Asher Trockman[1]
Alexander Robey[1] Tom Goldstein[2] Fei Fang[1] J. Zico Kolter[1]

[1]Carnegie Mellon University [2]University of Maryland

Contact us: yixuanx@cs.cmu.edu



Distillation *exists, and model owners are upset when their model gets distilled*

Meanwhile, companies accused of distillation also lack ways of self-certification

*Can we **rigorously** identify distillation?*

Ideally...

Performance preserves

Deployment is **cheap**

Outcome is **statistical**

Works for closed student

So that...

Model owners can deploy the method at a **small cost**

Claims of distillation can be made **accurately** without speculation

Red-and-green list watermark (Kirchenbauer et al. 2023)

For context $x_{1:l}$, define a **green list** computed as $S = H(x_{1:l}, k)$
H is a hash function keyed by k, outputing half of the vocabulary
Change the logits at sampling by increasing all green tokens by δ

$$z' = z + \delta \cdot S$$

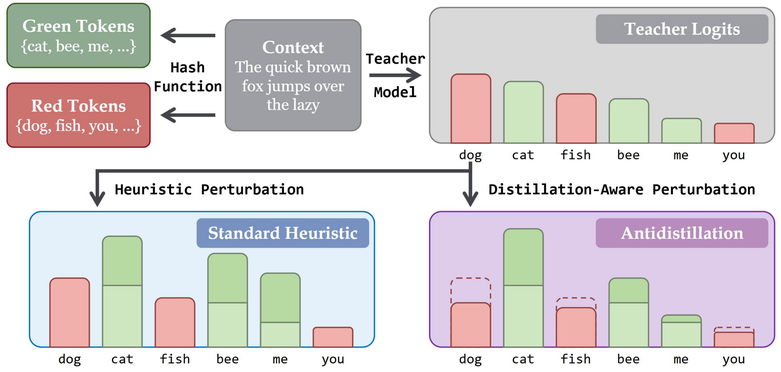
Normal texts have a **green-list ratio** of concentrated around 0.5
Watermarked texts' **green-list ratio** will be higher

Antidistillation fingerprinting

Antidistillation sampling (Savani et al. 2025) biases student loss function, adapt this technique to bias the **green-list ratio**
Omitting technical details, the final form looks like

$$z' = z + \lambda \cdot \Delta \text{ where } \Delta_t = \Pr[\text{The next token is } t] \cdot ([t \in S] - L)$$

Interpretation: Upweight perturbation for more likely tokens



Realistic Setting

Proxy ≠ Student
Student architecture unknown

Unsupervised Data
Fine-tuning data unknown

Closed-Weight Student
API access to student only

Multiple Domains

Math (GSM8K)

Conversation (OASST1)

Coding (MBPP)