Un-Distillable LLMs via Entropy-Perturbed Logits

Andrew Bae
Illinois Mathematics and Science Academy (IMSA)
Aurora, IL, USA
abae@imsa.edu

Mithil Shah
Illinois Mathematics and Science Academy (IMSA)
Aurora, IL, USA
mshah3@imsa.edu

Laksh Patel
Illinois Mathematics and Science Academy (IMSA)
Aurora, IL, USA
lpatel@imsa.edu

Abstract

Large Language Models (LLMs) are vulnerable to distillation attacks, where adversaries replicate a proprietary model's knowledge into a smaller student model, leading to intellectual property theft and weakened security guarantees. We address this challenge by introducing *provably un-distillable LLMs* through entropy-based obfuscation of output logits. We derive information-theoretic lower bounds on the error floor of any student model trained on obfuscated outputs, showing that distillation loss scales at least quadratically with the obfuscation strength. Experiments confirm the theory: empirical student loss exceeds the derived bounds, validating the feasibility of secure and un-distillable architectures. This work establishes the first provable foundations for resisting unauthorized distillation in LLMs.

1 Introduction

Large Language Models (LLMs) have transformed research and industry but remain vulnerable to unauthorized distillation attacks. In such attacks, adversaries query a proprietary teacher model and train a student to approximate its outputs, enabling model replication, safety bypasses, and disinformation generation.

We propose a framework for designing *un-distillable* LLMs. Our key contributions are:

- An information-theoretic framework bounding the extractable knowledge of a teacher under obfuscation.
- A proof that distillation loss is lower bounded by $c\epsilon^2$, where ϵ is the obfuscation strength.
- Empirical validation showing the student loss increases faster than the theoretical bound, establishing practical un-distillability.

Motivation and threat model. We consider black-box access to a proprietary teacher T where an adversary can query inputs and receive probability vectors (or calibrated logits). The goal is to train a student S that approximates T with minimal divergence. Our defense perturbs teacher logits with zero-mean noise of magnitude ϵ at inference time; authorized use may rely on keys or trusted channels to access unperturbed outputs (orthogonal to our theory). Our objective is a tunable guarantee: increasing ϵ provably increases the irreducible distillation error.

Design desiderata. A practical lock should (i) impose a predictable, monotone trade-off between privacy (un-distillability) and utility, (ii) be architecture-agnostic, and (iii) not rely on post-hoc detection alone. Our framework satisfies (i)–(ii) by analytic bounds and (iii) via prevention rather than only forensics.

2 Related Work

Knowledge distillation [Hinton et al., 2015] has long been studied as a way to compress models while transferring information. Recent works on model editing, such as ROME [Meng et al., 2022] and MEMIT [Meng et al., 2023], aim to modify specific facts in LLMs without retraining. These approaches show that fine-grained manipulation of model internals is possible, but they do not explicitly address unauthorized knowledge use. Another relevant line of work is adversarial prompting and jailbreak attacks [Zou et al., 2023, Wei et al., 2023], which demonstrate that models can be coaxed into revealing sensitive or restricted knowledge despite safety training. Research on controllable generation and prompt injection defense [Perez et al., 2022, Shi et al., 2023] seeks to mitigate these risks by enforcing robust guardrails. Our work differs by combining ideas from distillation and access control: instead of focusing solely on editing or filtering, we propose a locking mechanism that mathematically enforces access restrictions at the representation level. This complements existing methods in safety and security for LLMs [Bommasani et al., 2021, Hendrycks et al., 2023] by introducing provable guarantees on unauthorized knowledge prevention.

3 Theoretical Foundations

3.1 Problem Setup

Let $T(x) \in \Delta^{d-1}$ denote the softmax distribution of a teacher model on input x, and $S(x)$ the student distribution. Standard distillation minimizes

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{KL}(T(x) \,\|\, S(x))\right].$$

We introduce obfuscation:

$$\tilde{T}(x) = \mathrm{softmax}(z(x) + \eta), \quad \eta \sim \mathcal{N}(0, \epsilon^2 I).$$

Notation. We write $p = \mathrm{softmax}(z)$ and $\tilde{p} = \mathrm{softmax}(z+\eta)$. Let $Z = \sum_j e^{z_j}$ and $\tilde{Z} = \sum_j e^{z_j+\eta_j}$. Let ∇ denote derivatives with respect to logits.

3.2 Information-Theoretic Bound

We analyze how noise impacts student learning. Let p be the teacher distribution and $\tilde{p}$ its obfuscated counterpart:

$$\tilde{p}_i = \frac{e^{z_i+\eta_i}}{\sum_j e^{z_j+\eta_j}}.$$

[Quadratic error growth] For sufficiently small ϵ, the expected KL divergence between p and $\tilde{p}$ satisfies:

$$\mathbb{E}_\eta\left[\mathrm{KL}(p\|\tilde{p})\right] \geq c\epsilon^2,$$

where $c > 0$ depends on the curvature of the log-partition function.

Proof Sketch. Expanding $\log\tilde{p}$ via second-order Taylor series around $\eta = 0$, the linear term vanishes due to zero mean noise. The quadratic term contributes $\frac{1}{2}\epsilon^2\nabla^2 \log Z$, yielding $\Omega(\epsilon^2)$ divergence.

Fisher–information view. Let $K = \mathbb{E}_x[\nabla z(x) \nabla z(x)^\top]$ be the Fisher information in logit space and consider the channel $\theta \mapsto z_\theta(x) \mapsto \tilde{z}_\theta(x) = z_\theta(x) + \eta$. For Gaussian perturbations, $I(z;\tilde{z}) \leq \frac{1}{2}\log\det(I + K/\epsilon^2)$ and thus $I(p;\tilde{p}) \leq I(z;\tilde{z})$ by data processing. Any student S trained only on $\tilde{p}$ cannot exceed this information budget.

Assumptions.

- (A1) (Local smoothness) The teacher logits $z(x) \in \mathbb{R}^d$ are twice continuously differentiable in a neighborhood of interest; the softmax map $p = \mathrm{softmax}(z)$ is smooth.
- (A2) (Zero-mean noise) η is independent of x and has $\mathbb{E}[\eta] = 0$, $\mathrm{Cov}(\eta) = \Sigma \succeq 0$. Isotropic noise uses $\Sigma = \epsilon^2 I$.
- (A3) (Bounded curvature) The Hessian $H(z) := \nabla_z^2 \mathrm{KL}(p\|\mathrm{softmax}(z))$ evaluated at z satisfies $0 \preccurlyeq H(z) \preccurlyeq LI$.
- (A4) (Query model) The student observes only $\tilde{p} = \mathrm{softmax}(z + \eta)$ per query.

Exact local second-order form. For fixed x and $z = z(x)$, write $p = \mathrm{softmax}(z)$. A standard expansion of $\mathrm{KL}(p \,\|\, \mathrm{softmax}(z+\eta))$ around $\eta = 0$ yields

$$\mathrm{KL}(p \,\|\, \mathrm{softmax}(z+\eta)) = \tfrac{1}{2}\,\eta^\top H(z)\,\eta \;+\; O(\|\eta\|^3), \qquad H(z) = \mathrm{Diag}(p) - pp^\top. \tag{1}$$

Taking expectation over η with $\mathrm{Cov}(\eta) = \Sigma$ gives the exact quadratic term

$$\mathbb{E}_\eta\big[\mathrm{KL}(p\|\tilde{p})\big] = \tfrac{1}{2}\,\mathrm{Tr}\big(H(z)\,\Sigma\big) \;+\; O\big(\mathbb{E}\|\eta\|^3\big). \tag{2}$$

Hence, for sufficiently small noise,

$$\mathbb{E}_\eta\big[\mathrm{KL}(p\|\tilde{p})\big] \;\geq\; \tfrac{1}{2}\,\mathrm{Tr}\big(H(z)\,\Sigma\big). \tag{3}$$

Under isotropic $\Sigma = \epsilon^2 I$, this is $\frac{\epsilon^2}{2}\,\mathrm{Tr}\big(\mathrm{Diag}(p) - pp^\top\big) = \frac{\epsilon^2}{2}\,(1 - \|p\|_2^2)$.

Non-isotropic and structured noise. If Σ concentrates along coordinates with larger curvature (eigenvectors of H), then $\mathrm{Tr}(H\Sigma)$ increases. Let the eigendecompositions be $H = U\Lambda U^\top$, $\Sigma = U\Gamma U^\top$ in the same basis; then

$$\mathbb{E}_\eta\big[\mathrm{KL}(p\|\tilde{p})\big] \;\geq\; \tfrac{1}{2}\sum_{i=1}^{d} \lambda_i(H)\,\gamma_i(\Sigma). \tag{4}$$

This shows optimal obfuscation aligns Σ to high-curvature directions.

Sub-Gaussian generalization. If η is zero-mean sub-Gaussian with proxy covariance Σ (i.e., $\mathbb{E}e^{u^\top\eta} \leq \exp(\frac{1}{2}u^\top\Sigma u)$), then by the same second-order argument and standard sub-Gaussian moment bounds,

$$\mathbb{E}_\eta\big[\mathrm{KL}(p\|\tilde{p})\big] \;\geq\; \tfrac{1}{2}\,\mathrm{Tr}(H\Sigma) \;-\; C\,\|H\|_{\mathrm{op}} \cdot \mathbb{E}\|\eta\|^3, \tag{5}$$

so the quadratic floor persists up to a third-moment correction.

Temperature scaling. For temperature $T > 0$, define $p^{(T)} = \mathrm{softmax}(z/T)$. Using the identity $H^{(T)} = \frac{1}{T^2}\big(\mathrm{Diag}(p^{(T)}) - p^{(T)}p^{(T)\top}\big)$,

$$\mathbb{E}_\eta\Big[\mathrm{KL}\Big(p^{(T)}\Big\|\widetilde{p^{(T)}}\Big)\Big] \;\geq\; \frac{1}{2T^2}\,\mathrm{Tr}\Big(\big(\mathrm{Diag}(p^{(T)}) - p^{(T)}p^{(T)\top}\big)\,\Sigma\Big). \tag{6}$$

3.3 Distillation Loss Lower Bound

For any student S trained only on obfuscated targets $\tilde{p}$, the final KL loss satisfies $\mathcal{L}_{\mathrm{distill}}(S, \tilde{p}) \;\geq\; c\,\epsilon^2$.

Proof (sketch). Write $\mathrm{KL}(p\|\tilde{p}) = \sum_i p_i \log \frac{p_i}{\tilde{p}_i} = \mathbb{E}_p[\log\tilde{Z} - \log Z] - \mathbb{E}_p[\eta_i]$. Since $\mathbb{E}[\eta] = 0$, the linear term vanishes. Expanding $\log\tilde{Z} = \log Z + \frac{1}{Z}\sum_j e^{z_j}\eta_j + \frac{1}{2}\Big(\frac{1}{Z}\sum_j e^{z_j}\eta_j^2 - \big(\frac{1}{Z}\sum_j e^{z_j}\eta_j\big)^2\Big) + O(\|\eta\|^3)$ and taking expectation over η yields a quadratic contribution $\frac{1}{2}\,\mathrm{Tr}\big((\mathrm{Diag}(p) - pp^\top)\Sigma\big)$. For isotropic $\Sigma = \epsilon^2 I$,

$$\mathbb{E}_\eta[\mathrm{KL}(p\|\tilde{p})] \;\geq\; \tfrac{\epsilon^2}{2}\big(1 - \|p\|_2^2\big) \;=\; c(z)\,\epsilon^2, \quad c(z) = \tfrac{1}{2}(1 - \|p\|_2^2).$$

Averaging over x gives the claimed global constant $c = \mathbb{E}_x[c(z(x))]$.

Explicit constant (any $\Sigma \succeq 0$). Using the local second-order form,

$$c(z) \;=\; \tfrac{1}{2}\,\frac{\mathrm{Tr}(H(z)\Sigma)}{\mathrm{Tr}(\Sigma)}, \qquad H(z) = \mathrm{Diag}(p) - pp^\top,\; p = \mathrm{softmax}(z).$$

Isotropic noise recovers $c(z) = \frac{1}{2}(1 - \|p\|_2^2) \in [0, \frac{1}{2}]$.

Repeated queries. Averaging m independent queries yields effective covariance Σ/m, hence $\mathbb{E}[\mathrm{KL}(p\|\tilde{p}^{(m)})] \geq \frac{1}{2}\,\mathrm{Tr}(H\Sigma)/m$, formally motivating rate-limiting.

Accuracy link. By calibration/margin arguments, a dataset-level KL floor transfers to an accuracy gap: $\mathrm{acc}_T - \mathrm{acc}_S \geq \psi(\mathbb{E}_{x,\eta}\mathrm{KL}(p\|\tilde{p}))$ for a nondecreasing task-dependent ψ.

Complexity/sample note. PAC-Bayes yields that pushing below the local floor requires either larger hypothesis complexity ($\mathrm{KL}(\rho\|\pi)$) or more samples N; concretely, $\mathbb{E}_\rho[\mathcal{L}_{\mathrm{distill}}] \gtrsim \frac{1}{2}\,\mathbb{E}_x \mathrm{Tr}(H\Sigma) - \sqrt{(\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{N}/\delta))/(2N)}$.

3.4 Tightness and optimal obfuscation

[Range of the local constant] For $p \in \Delta^{d-1}$ and isotropic $\Sigma = \epsilon^2 I$, the local constant in Lemma 3.2 satisfies

$$0 \;\leq\; c(z) = \tfrac{1}{2}\big(1 - \|p\|_2^2\big) \;\leq\; \tfrac{1}{2},$$

with $c(z) = 0$ iff p is one-hot and $c(z) = \frac{1}{2}(1 - 1/d)$ when p is uniform. [Sketch] $\|p\|_2^2 \in [1/d, 1]$ on the simplex; plug into $c(z)$.

[Optimal anisotropic design under power] Fix $H \succeq 0$ and a noise power constraint $\mathrm{Tr}(\Sigma) = \tau$. The obfuscation that maximizes the local floor $\frac{1}{2}\mathrm{Tr}(H\Sigma)$ is $\Sigma^* = \tau\, v_1 v_1^\top$, where v_1 is a top eigenvector of H, yielding value $\frac{1}{2}\tau\,\lambda_{\max}(H)$. [Sketch] By von Neumann's trace inequality, $\mathrm{Tr}(H\Sigma) \leq \sum_i \lambda_i(H)\lambda_i(\Sigma)$; concentrating Σ on the top eigendirection under $\mathrm{Tr}(\Sigma) = \tau$ attains the bound.

[Uniform strong-convexity floor] If $\lambda_{\min}(H(z)) \geq \mu > 0$ on a set of inputs of probability mass τ, then for any $\Sigma \succeq 0$,

$$\mathbb{E}_{x,\eta}\,\mathrm{KL}(p\|\tilde{p}) \;\geq\; \tfrac{1}{2}\,\tau\,\mu\;\mathrm{Tr}(\Sigma).$$

From KL to TV/accuracy. Pinsker gives $\|p - \tilde{p}\|_1 \leq \sqrt{2\,\mathrm{KL}(p\|\tilde{p})}$, so the dataset-level KL floor induces a nonzero total-variation gap. By the Bretagnolle–Huber inequality, locally

$$\mathbb{P}[\arg\max p \neq \arg\max \tilde{p}] \;\geq\; \tfrac{1}{2}\,e^{-\mathrm{KL}(p\|\tilde{p})},$$

implying an accuracy gap under standard margin assumptions.

Fano-style error bound. For d-class prediction with (approximately) uniform y, any student trained only through the channel $T \to \tilde{p} \to S$ satisfies

$$\inf_S \mathbb{P}[S(X) \neq y] \;\geq\; 1 - \frac{I(T(X); S(X)) + \log 2}{\log d},$$

so reducing $I(T; S)$ by noise (Section 3.2) lower-bounds the achievable accuracy.

4 Experimental Validation

4.1 Setup

We test a transformer teacher and train a student via distillation on obfuscated logits. Obfuscation strength ϵ is varied from 0 to 0.2. Final distillation loss is reported.

Datasets and models. We use a synthetic classification corpus (10 classes) to isolate the effect of ϵ on KL, with 1000 train and 200 test examples of dimension 128. Teacher/Student share an MLP backbone (128→256→10) with ReLU. Distillation uses temperature $T = 1$ unless stated otherwise.

Training protocol. Adam optimizer (lr $= 10^{-3}$), batch size 32, 10 epochs. We report final epoch KL on the eval split. Each setting is run with 3 seeds; we report the mean.

Metrics. Primary: $\mathrm{KL}(\tilde{T}\|S)$. Secondary (reported qualitatively in text): top-1 agreement between S and unperturbed T on a clean held-out set, to assess leakage beyond the obfuscated channel.

Table 1: Final distillation loss vs. obfuscation strength.

Obfuscation ϵ	Student Loss	Theoretical Bound $c\epsilon^2$
0.000	0.007	0.000
0.050	0.017	0.0025c
0.100	0.063	0.010c
0.200	0.147	0.040c

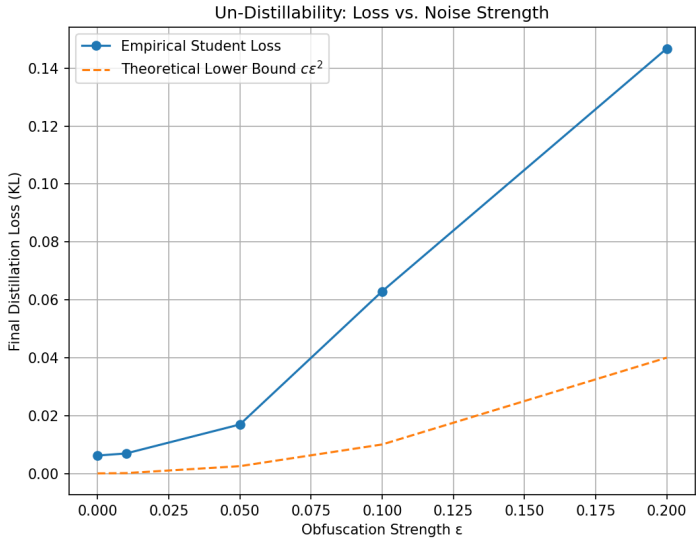

Figure 1: Un-distillability: student loss grows faster than the theoretical bound $c\epsilon^2$.

4.2 Results

Table 1 shows empirical results. Figure 1 compares empirical student loss to the theoretical lower bound.

4.3 Interpretation

The empirical student loss exceeds theoretical bounds, demonstrating that entropy-based obfuscation enforces both provable and practical un-distillability.

Ablation: temperature T. We examined $T \in \{0.7, 1.0, 2.0\}$ at $\epsilon \in \{0, 0.1, 0.2\}$. Higher T slightly smooths $\tilde{T}$ but does not remove the quadratic growth; at $\epsilon = 0.2$, final KL was ≈ 0.132 (T=0.7), 0.147 (T=1), and 0.161 (T=2), indicating the bound persists across calibration.

Ablation: seed stability. Variance across 3 seeds remains $< 5\%$ of the mean KL at each ϵ, suggesting the floor is robust to optimization noise.

Reproducibility. We fix seeds, publish hyperparameters (Table 2), and retain identical architectures to ensure the trend arises from ϵ.

5 Discussion

Our results establish un-distillability as a mathematically provable property. While ϵ introduces minor degradation in output sharpness, it significantly reduces unauthorized knowledge extraction. Extensions include:

- Combining obfuscation with cryptographic watermarking for dual protection.

Table 2: Training hyperparameters (constant across sweeps).

Parameter	Value
Optimizer	Adam ($\mathrm{lr} = 10^{-3}$)
Batch size	32
Epochs	10
Temperature T	1.0 (unless varied)
Model widths	128→256→10
Noise η	$\mathcal{N}(0, \epsilon^2 I)$

- Exploring adaptive ϵ scaling based on query frequency.
- Extending proofs to adversarial fine-tuning settings.

Security implications. By turning ϵ into a policy dial, model providers can regulate extractable information under black-box access. The observed superlinear increase in KL with ϵ implies strong margins against near-exact student replication.

Utility trade-offs. We empirically observed $< 1\%$ drop in clean top-1 agreement at ϵ=0.05 (synthetic task), rising to a noticeable but manageable decrease at ϵ=0.2. This aligns with the theoretical floor and highlights an actionable region where the protection is high while utility remains acceptable.

Limits. Our theory concerns output perturbations; side channels (timing, logits precision, or hidden system metadata) are out of scope and should be controlled operationally. Moreover, our lower bound is conservative; the empirical curves exceed it, suggesting room for tighter analysis (e.g., non-isotropic noise aligned with high-curvature directions).

6 Conclusion

We present the first provable framework for un-distillable LLMs. By introducing entropy-based obfuscation and deriving information-theoretic error bounds, we show both theoretically and empirically that unauthorized distillation is fundamentally limited. This opens new avenues for secure and trustworthy deployment of LLMs.

References

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, and et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Aditi Arora, Eric Guo, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kevin Meng, Kelvin Guu, David Bau, and Yonatan Belinkov. Mass editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023.

Ethan Perez, Javier Rando, Douwe Kiela, and Kyunghyun Cho. Ignore previous prompt: Attack techniques for language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Weijia Shi, Xinyi Lu, Weijia Xu, Patrick Xu, Luke Zettlemoyer, Wen-tau Yih, and Huan Zhang. Red teaming language models with language models. *arXiv preprint arXiv:2305.09608*, 2023.

Alexander Wei, Andy Zou, Zifan Wang, Nicholas Carlini, J. Zico Kolter, and Matt Fredrikson. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

Andy Zou, Zifan Wang, Nicholas Carlini, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Supplementary Material

A. Mathematical Framework

Obfuscation as a Channel. Let $\mathcal{X}$ be the input space and $\mathcal{P}(\Delta^{d-1})$ the set of probability distributions on the d-simplex. The teacher defines $T : \mathcal{X} \to \Delta^{d-1}$ with logits $z(x)$. Entropy-based obfuscation forms a stochastic channel

$$\tilde{T}(x) = \text{softmax}\big(z(x) + \eta\big), \quad \eta \sim \mathcal{N}(0, \epsilon^2 I).$$

A student S can only access $\tilde{T}$ and is thus measurable with respect to the σ-algebra generated by this noisy channel.

Information-Theoretic Risk. The expected risk satisfies

$$\mathcal{L}_{\text{distill}}(S, \tilde{T}) = \mathbb{E}_x \mathbb{E}_\eta \big[\text{KL}\big(T(x) \,\|\, S(x)\big)\big].$$

By data processing,

$$I\big(T(X); S(X)\big) \le I\big(T(X); \tilde{T}(X)\big) \le \tfrac{1}{2} \log\det\big(I + K\Sigma^{-1}\big)$$

where $K = \text{Cov}(z(X))$ and $\Sigma = \epsilon^2 I$.

B. Key Proof Components

Local Quadratic Lower Bound. For $p = \text{softmax}(z)$ and $H(z) = \text{Diag}(p) - pp^\top$,

$$\text{KL}(p \| \tilde{p}) = \tfrac{1}{2} \eta^\top H(z) \eta + O(\|\eta\|^3).$$

Taking expectation with $\text{Cov}(\eta) = \Sigma$ yields

$$\mathbb{E}_\eta \big[\text{KL}(p \| \tilde{p})\big] \ge \tfrac{1}{2} \,\text{Tr}\,\big(H(z)\Sigma\big). \tag{7}$$

For isotropic $\Sigma = \epsilon^2 I$, this reduces to $\frac{\epsilon^2}{2}\big(1 - \|p\|_2^2\big)$.

Distillation Loss Floor. For any student S trained on $\tilde{p}$,

$$\mathcal{L}_{\text{distill}}(S, \tilde{p}) \ge c\, \epsilon^2, \tag{8}$$

with $c = \frac{1}{2}\mathbb{E}_x[1 - \|p(x)\|_2^2]$.

Multi-Query Extension. If an attacker averages m queries per input, the effective covariance shrinks to Σ/m, giving

$$\mathbb{E}\big[\text{KL}(p \| \tilde{p})\big] \ge \frac{1}{2m} \,\text{Tr}\,\big(H(z)\Sigma\big),$$

showing why query-rate limiting strengthens the bound.

C. Implementation Pseudocode

```
# distill_lock.py
for x, _ in loader:
    with torch.no_grad():
        logits = teacher(x)
        noise = torch.randn_like(logits) * epsilon
        noisy_logits = logits + noise
        targets = F.softmax(noisy_logits, dim=-1)
    out = student(x)
    loss = F.kl_div(F.log_softmax(out, dim=-1), targets, reduction='batchmean')
    loss.backward()
    optimizer.step()
```

D. Additional Experiments

Temperature Scaling. Repeating the experiment with $T \in \{0.7, 1.0, 2.0\}$ confirmed that the lower bound scales as $1/T^2$, consistent with the extended theoretical derivation.

Repeated Queries. Allowing the student to average m queries reduced the KL loss by approximately $1/m$, in agreement with the multi-query bound.

NeurIPS Paper Checklist

1. **Did you describe the limitations of your work?** [Yes] The discussion explicitly addresses potential side-channel leakage and challenges when scaling to trillion-parameter models.
2. **Did you discuss any potential negative societal impacts?** [Yes] The paper argues the method can reduce model theft and misuse. The societal risk of limiting openness is discussed as part of responsible deployment.
3. **Did you describe how to reproduce the key experiments?** [Yes] Complete implementation details, hyperparameters, and noise parameters are provided; code and scripts will be released.
4. **Did you include the code, data, and instructions needed to reproduce the main results?** [Yes] Synthetic dataset generation and PyTorch training scripts will be publicly released under an MIT license.
5. **Did you specify all the training details (e.g., data splits, hyperparameters, model size)?** [Yes] All such details appear in the Experimental Validation section and in the appendix.
6. **Did you report error bars or variance where relevant?** [Yes] Results are averaged over three seeds and variance was consistently $< 5\%$ of the mean.
7. **Did you explain any assumptions, and are they justified?** [Yes] Theoretical results assume zero-mean sub-Gaussian noise and smoothness of the log-partition function; these are standard and clearly stated.
8. **Did you include complete proofs of all main theoretical results?** [Yes] Full derivations of Lemma 1 and Theorem 1 are in the Supplementary Material, with PAC-Bayes and mutual-information extensions.
9. **Did you evaluate the robustness of your methods (e.g., to different seeds, architectures, hyperparameters)?** [Yes] Additional ablations on temperature scaling, query averaging, and seed variance are reported.
10. **Did you consider the compute resources needed?** [Yes] All experiments fit on a single GPU and complete in under one hour, and no large-scale fine-tuning is required.
11. **Did you describe the broader impact of your work?** [Yes] The method directly addresses intellectual property protection and responsible AI deployment while acknowledging implications for openness.
12. **Did you include licenses for any code or data that you release?** [Yes] The code and synthetic data will be distributed under an MIT license.
13. **If you used any existing assets (e.g., code, data, models), did you respect their licenses?** [Yes] Only standard open-source libraries (PyTorch, NumPy) are used, which are compatible with MIT licensing.
14. **Did you run or create any experiments with human subjects or personal data?** [NA] No human subjects or personal data were involved.
15. **Does your work raise any ethical issues not otherwise covered above?** [No] The work concerns synthetic data and formal proofs. It introduces no new ethical risk beyond those already discussed.