SMI: Statistical Membership Inference for Reliable Unlearned Model Auditing

NeurIPS 2026
Machine Unlearning
Statistical Auditing

Jialong Sun[1*] Zeming Wei[2*] Jiaxuan Zou[3*] Jiacheng Gong[1] Jie Fu[4] Chengyang Dong[5] Heng Xu[5] Jialong Li[5] Bo Liu[5†]

[1] School of Cybersecurity, Heilongjiang University · [2] School of Mathematical Sciences, Peking University · [3] School of Mathematics and Statistics, Xi'an Jiaotong University
[4] Department of Computer Science, Stevens Institute of Technology · [5] Faculty of Computer Science and Artificial Intelligence, Shenzhen University of Advanced Technology

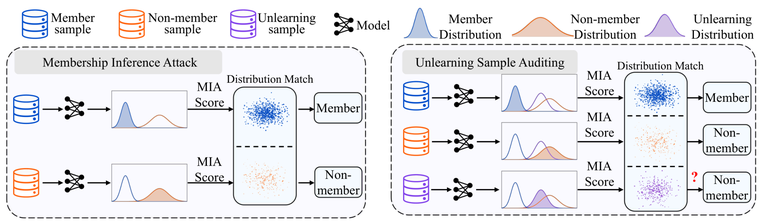
KEY IDEA

Failed membership inference does not imply true forgetting.

SMI estimates the non-member-like proportion of the audited distribution, without training shadow models.

01 Why attacks can mislead

MIA learns member/non-member separation. Unlearning shifts the audit distribution.



The mismatch: an "undetected" sample need not behave like a genuine non-member. [Fig. 1]

An additional auditing-error term

PAC-Bayes setting: bounded loss, i.i.d. attacker training data, fixed prior P and posterior Q ≪ P. With probability at least 1 − δ: [Cor. 2.2]

$$R_{\mathcal{F}_u}(Q) \le R_S(Q) + \varepsilon_{\text{stat}} + \varepsilon_{\text{shift}}$$

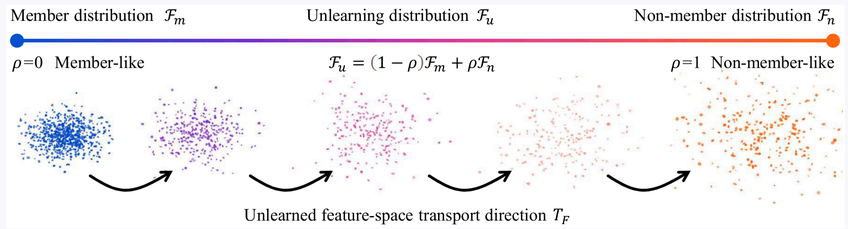
$R_S(Q)$	Empirical risk on attack data
$\varepsilon_{\text{stat}} = \sqrt{\frac{2[\chi^2(Q\|P)+1]}{m\delta}}$	Statistical complexity
$\varepsilon_{\text{shift}} = \sqrt{\frac{1}{2}D_\infty(\mathcal{F}_t\|\mathcal{F}_u)}$	Distribution shift

Better attack training alone does not certify forgetting. The bound also depends on the shift from attack-training data to audited data.

D_∞: order-∞ Rényi divergence; m: sample size.
δ: failure probability.

02 Statistical membership inference

A feature-space transport perspective. Approximate the audited distribution as a mixture of two references:



$$\mathcal{F}_u \approx (1-\rho)\mathcal{F}_m + \rho\mathcal{F}_n, \quad \rho \in [0,1]$$

$\mathcal{F}_m$: member · $\mathcal{F}_n$: non-member · $\mathcal{F}_u$: audited

SMI: covariance matching

PROP. 3.1 / LEMMA 3.2

Mixture covariance, with $d = \mu_n - \mu_m$:

$$\Sigma(\rho) = (1-\rho)\Sigma_m + \rho\Sigma_n + \rho(1-\rho)dd^\top$$

$$\rho^* = \arg\min_{0\le\rho\le1} \|\Sigma_u - \Sigma(\rho)\|_F^2$$

Solve a cubic: check feasible stationary roots and endpoints.

SMI-M: kernel mean matching

LEMMA 3.3

Compare kernel mean embeddings μ^k in a Hilbert space.

$$\rho^* = \arg\min_{0\le\rho\le1} \left\|\mu_u^k - (1-\rho)\mu_m^k - \rho\mu_n^k\right\|_{\mathcal{H}}^2$$

Convex quadratic in ρ. Use logit gradients and an RBF kernel with median bandwidth in the experiments.

Quantify finite-sample sensitivity

Bootstrap the audit set, refit ρ, and report the **5th–95th percentile reference range**. Point estimate unchanged. [§3.3]

03 Evidence at a glance

Full retraining: SMI-M leads all six main-table settings. Ideal target: ρ* = 100%.

Model · data removed	Best MIA	SMI-M ρ*
R18 · C10 · 5%	88.67%	**92.44%**
R18 · C100 · 5%	89.36%	**89.84%**
R50 · C100 · 5%	87.04%	**92.64%**
R50 · C100 · 10%	87.38%	**93.95%**
R50 · CINIC · 10%	90.47%	**92.27%**
ViT · CINIC · 10%	89.84%	**93.62%**

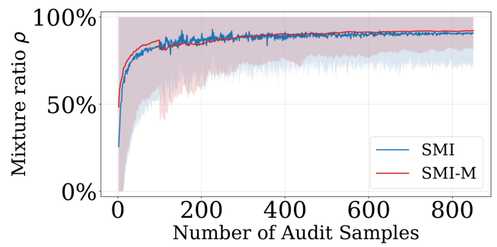
[Table 1] Best MIA: maximum per row. R18/R50: ResNet; C10/C100: CIFAR; CINIC: CINIC-10.

No shadow-model training

SMI: 0.93 s · SMI-M: 4.30 s
200 bootstrap iterations.
One shadow model training: **2,460 s.**

[Table 2, CIFAR-100] Excludes the separately reported 44.7 s inference pass.

Sample size controls stability



Stable near 200 samples. CINIC-10 / ResNet-50. Bands: reference ranges. [Fig. 3]

Measure how non-member-like the audited distribution becomes.

jialong.sunh@gmail.com · liubo@suat-sz.edu.cn
* Equal contribution † Corresponding author