SMI: Statistical Membership Inference for Reliable Unlearned Model 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
jialong.sunh@gmail.com liubo@suat-sz.edu.cn

Abstract

Machine unlearning (MU) is essential for enforcing the right to be forgotten in machine learning systems. A key challenge of MU is how to reliably audit whether a model has truly forgotten specified training data. Membership Inference Attacks (MIAs) are widely used for unlearned model auditing, where samples that evade membership detection are regarded as successfully forgotten. We show this assumption is fundamentally flawed: failed membership inference does not imply true forgetting. We prove that unlearned samples occupy a fundamentally different positions in the feature space feature space than non-member samples, making this alignment bias unavoidable and unobservable, which leads to systematically optimistic evaluations of unlearning performance. Meanwhile, training shadow models for MIA incurs substantial computational overhead. To address both limitations, we propose Statistical Membership Inference (SMI), a training-free auditing framework that reformulates auditing as estimating the non-member mixture proportion in the unlearned feature distribution. Beyond estimating the forgetting rate, SMI also provides bootstrap reference ranges for quantified auditing reliability. Extensive experiments show that SMI consistently outperforms all MIA-based baselines, with no shadow model training required. Overall, SMI establishes a principled and efficient alternative to MIA-based auditing methods, with both theoretical guarantees and strong empirical performance.

1 Introduction

Machine learning (ML) has been widely deployed in various critical domains, such as healthcare, finance, and Transportation, whose training processes may heavily depend on privacy-sensitive data under certain protections Fu et al. [2025], Liu et al. [2026]. However, the authentication of data usage from users is not necessarily unwarranted, particularly in the background of policies like GDPR that propose "the right to be forgotten", which indicates users reserve the right to withdraw the data authentication with considerations like privacy-protection, ethical issues, or policy requirements Le-Khac and Truong [2025], Nguyen et al. [2025], Liu et al. [2022b], Hayes et al. [2025]. This explicitly indicates an urgent need for an accountable mechanism that can timely remove the impact of related data on model behaviors when authentication is revoked, as well as restrictions on training, inference, or communication with these data for future model deployments Ye et al. [2025].

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Given this background, machine unlearning has emerged as a prominent approach to address these requirements Wei et al. [2025], Ebrahimpour-Boroojeny et al. [2025], Wu et al. [2025]. Generally, machine unlearning involves using post-training methods to eliminate the influence of specific training data on the trained model. Despite notable success, a key open question in machine unlearning remains: *How can we effectively evaluate the forgetting of a given set of data Zhang et al. [2025], Zheng et al. [2025]?* In this paper, we refer to this evaluation as the *auditing* problem of machine unlearning. Currently, common auditing criteria can be categorized into three types of backdoor attacks Gu et al. [2019], Liu et al. [2022a], adversarial attacks Szegedy et al. [2013], Nicolazzo et al. [2025], and membership inference attacks (MIA) Shokri et al. [2017], Han et al. [2024].

Both backdoor and adversarial attacks attempt to poison the training data, leading to compromised model performance on specific tasks. Since machine unlearning can be viewed as recovering the model's ability by locating the poisoning data and unlearning them, these two kinds of attacks calculate the attack success rate on the unlearned model for auditing Liu et al. [2025]. Thus, a critical concern emerges: as defense techniques (e.g., adversarial training against adversarial attacks) can also decrease ASRs, yet do not achieve the unlearned goal, the trustworthiness of such auditing methods is compromised. In addition, both attack strategies require training a new target model, which may be perceived as tampering with the underlying task.

Besides, MIA aims to judge whether a specific data point is involved in the model training process, determined by assessing the output of the model given this input Wang et al. [2025b]. Intuitively, MIA can be mathematically formulated as a threshold-based classification problem:

$$\text{MIA}(x; w) = \mathbb{1}_{\text{scoreMIA}(x;w)\geq\beta}, \tag{1}$$

where $\text{scoreMIA}(x; w)$ is the membership score for sample x under model parameter w. So far, a thread of MIA-based methods have pave the way for effectively auditing machine unlearning results Nasr et al. [2019]. Many variants have been implemented to improve MIA performance on diverse unlearning settings, such as RMIA Zarifzadeh et al. [2024] , IAM Wang et al. [2025a] and RULI Naderloui et al. [2025]. These attack-based auditing methods are prone to a failure mode in which failed attacks give rise to illusory forgetting. Specifically, a failure in the attack strategy may be misinterpreted as the model having unlearned the sample. Consequently, this causes the auditing of the unlearning task to fail.

In this study, we aim to address the issue of illusory forgetting caused by failed membership inference attacks. We revisit the validity of the MIA paradigm. In an MIA task, the goal is essentially to train a binary classifier that distinguishes member sample features F_m (e.g., features from the training dataset) from non-member sample features F_n (e.g., features from the test dataset). However, in a real unlearning task, the audited targets are the feature representations of unlearned samples F_u Du et al. [2025], Bai et al. [2025], Miyamoto et al. [2025]. Conventional MIA assumes that the audit features of F_u are close to those of F_n, and therefore treats samples whose features in F_u are classified as F_n as successfully unlearned. As illustrated in Figure 1, this assumption is not well aligned with the actual objective of unlearning. This is because, in the unlearning task, F_m corresponds to feature representations that contribute positive gradients to the model, F_n corresponds to feature representations that contribute no gradient, while F_u corresponds to feature representations that induce negative gradients. Such a gradient-based distinction is fundamentally misaligned with the core assumption of MIA, which does not explicitly account for these opposing optimization dynamics. Although some studies have attempted to improve the alignment between MIA and unlearning through group-level attack strategies or independent modeling procedures, these methods do not fundamentally resolve the alignment bias and still require training tens of shadow models, each as costly as the audited model itself, making the computational overhead prohibitive.

Based on the above observations, we first conduct a theoretical analysis from two aspects. First, we formally prove that MIA-based strategies for auditing machine unlearning suffer from inevitable alignment errors. Second, based on the sample features migration property during the unlearning process, we introduce optimal transport theory to derive the corresponding optimal estimation. Building on this analysis, we propose a training-free statistical method for auditing unlearning tasks, called Statistical Membership Inference (SMI). Unlike existing methods that rely on shadow model training and classifier training, SMI does not require any shadow models or MIA classifiers. Instead, it only uses statistical techniques to estimate the mixture proportion between member and non-member distributions. In addition, SMI provides reference ranges for the audit results, which offers a reliable reference for audit decisions. Overall, our contributions are summarized as follows:

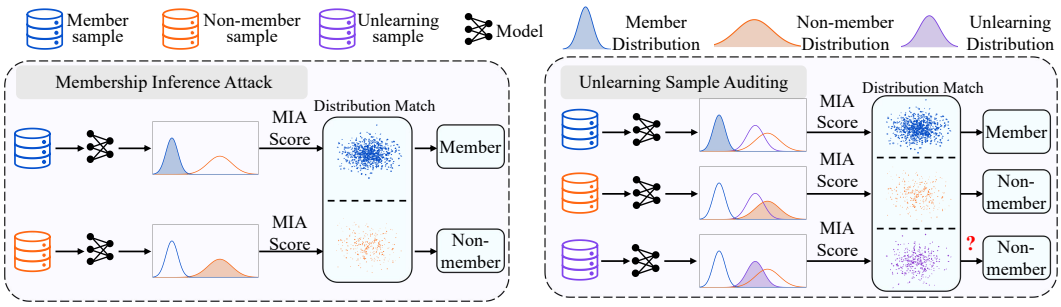


Figure 1: Misalignment Between MIA and the Unlearned Model Auditing Process

- **Conceptually**, we show that failed MIA does not imply true forgetting, since unlearned samples may reside in a feature distribution fundamentally distinct from that of genuine non-members. This reveals a structural mismatch between MIA-based auditing and the goal of unlearned model auditing.
- **Theoretically**, we prove that this mismatch arises from an unavoidable yet unobservable alignment bias between unlearned samples and non-members. We further reformulate unlearned model auditing as estimating the member and non-member mixture proportion, shifting the Monge problem of unlearning from the parameter space to the feature space.
- **Practically**, we propose SMI and its variant SMI-M, a training-free auditing framework that estimates the forgetting rate without shadow models or learned classifiers, while providing bootstrap reference ranges to quantify auditing reliability.

2 Revisiting MIA: From Attack to Auditing

This section explains the limitations of MIA in unlearned model auditing and motivates the proposed improvement, SMI. First, we use PAC-Bayes theory to derive the alignment bias in unlearned model auditing. Second, we adopt an optimal-transport perspective to revise the auditing strategy.

Alignment Bias in MIA. We decompose the attacker's learning process to identify the alignment bias in MIA-based unlearned model auditing. Let F_n, F_m, and F_u be the auditing feature sets of the retained member dataset D_n, non-member dataset D_m, and unlearning dataset D_u, and let $\mathcal{F}_n$, $\mathcal{F}_m$, and $\mathcal{F}_u$ be the corresponding empirical distributions Chen et al. [2021], Luo et al. [2025]. Let $\mathcal{P}$ and $\mathcal{Q}$ denote the prior and posterior distributions over the attacker's hypothesis space Honorio and Jaakkola [2014], Ohnishi and Honorio [2021]. The χ^2-based PAC-Bayes risk bound is stated in Theorem 2.1.

Theorem 2.1 (PAC-Bayesian Bound for MIA)**.** *For a binary classifier, suppose that the attacker model $\mathcal{A}$ learns from a prior distribution $\mathcal{P}$ and induces a posterior distribution $\mathcal{Q}$. For a sample size m, with probability at least $1-\delta$, the following inequality holds for any distribution $\mathcal{Q}$:*

$$R_D(\mathcal{Q}) \leq \underbrace{R_S(\mathcal{Q})}_{\textit{Empirical Risk}} + \underbrace{\sqrt{\chi^2(\mathcal{Q}\|\mathcal{P})+1}\cdot\sqrt{\frac{2}{\delta m}}}_{\textit{Statistical Error}}, \tag{2}$$

where $R_D(\mathcal{Q})$ is the true risk when hypotheses are sampled according to $\mathcal{Q}$, $R_S(\mathcal{Q})$ is the average empirical risk computed on the training set, and $\chi^2(\mathcal{Q}\|\mathcal{P})$ is the χ^2 divergence between $\mathcal{Q}$ and $\mathcal{P}$.

Remark 1: We use a χ^2-based PAC-Bayes framework because it provides convenient second-moment control for deriving the subsequent distribution-shift term. Detailed proof refers to Appendix C.1.

Theorem 2.1 gives a general decomposition of the attacker's learning error. More importantly, it does not reveal the alignment discrepancy in unlearned model auditing. The reason is that the bound is derived on the MIA training distribution $\mathcal{F}_t$, which is constructed from $\mathcal{F}_n$ and $\mathcal{F}_m$, while the actual auditing target is $\mathcal{F}_u$. Conventional MIA implicitly treats $\mathcal{F}_t$ and $\mathcal{F}_u$ as aligned, which can introduce an alignment bias. We therefore incorporate the discrepancy between $\mathcal{F}_t$ and $\mathcal{F}_u$ into the PAC-Bayes bound and obtain a refined auditing-error decomposition in Corollary 2.2.

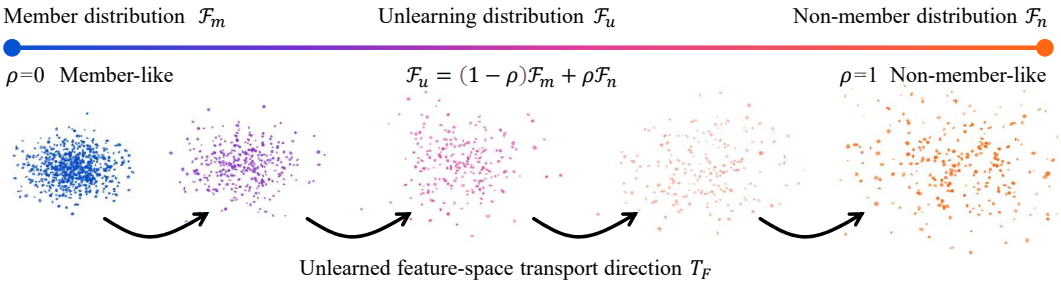


Figure 2: Approximate Representation of the Distribution Transport Problem as a Distribution Mixture Problem

Corollary 2.2 (PAC-Bayesian Bound for MIA Auditing Results)**.** *When the training data distribution $\mathcal{F}_t$ of the MIA attacker model $\mathcal{A}$ is used to audit the data distribution $\mathcal{F}_u$, for a sample size m, with probability at least $1-\delta$, the following risk inequality holds:*

$$R_{\mathcal{F}_u}(\mathcal{Q}) \leq R_S(\mathcal{Q}) + \underbrace{\sqrt{\frac{2}{m\delta}\left(\chi^2(\mathcal{Q}\|\mathcal{P})+1\right)}}_{\textit{Statistical Error}} + \underbrace{\sqrt{\frac{1}{2}D_\infty(\mathcal{F}_t\|\mathcal{F}_u)}}_{\textit{Auditing Error}}. \tag{3}$$

where $D_\infty(\mathcal{F}_t\|\mathcal{F}_u) = \log\left(\sup \frac{\mathcal{F}_t}{\mathcal{F}_u}\right)$ is the Rényi divergence Khamaru et al. [2025]. Please refer to Appendix C.2 for a detailed explanation of the notation and all proofs.

This corollary shows that the MIA risk bound contains an additional auditing term controlled by $D_\infty(\mathcal{F}_t\|\mathcal{F}_u)$. This term increases with the distribution shift between $\mathcal{F}_t$ and $\mathcal{F}_u$, indicating that conventional MIA faces an alignment bias in unlearned model auditing. Thus, the auditing error is not only caused by empirical risk or statistical complexity, but also by the mismatch between the attacker's training distribution and the actual auditing distribution.

Revising the Auditing Strategy. To address the alignment bias in conventional MIA, we revise the auditing criterion from the goal of machine unlearning. After sufficient unlearning, the post-unlearning model should no longer preserve membership traces of the unlearned samples. Thus, auditing should not only test whether the attacker fails to identify their membership. It should also examine whether their auditing features are close to those of non-member samples.

From a transport perspective, machine unlearning can be viewed as moving the original model state toward the retrained model state. Let α_w be the parameter distribution after original training, and let $\beta_{w\setminus u}$ be the ideal retraining distribution after removing D_u. The ideal unlearning objective can then be written as the following Monge-type problem.

Definition 2.3. The Monge problem in parameter space is:

$$\min_{T_w : T_{w\#}\alpha_w = \beta_{w\setminus u}} \int c_w\big(w, T_w(w)\big)\, d\alpha_w(w), \tag{4}$$

where T_w is the unlearning map in the parameter space, $T_{w\#}\alpha_w$ denotes the pushforward distribution obtained by transporting the original parameter distribution α_w through T_w, and c_w is the parameter transport cost. This objective requires the post-unlearning model distribution to be aligned with the ideal retrained model distribution.

Directly verifying this parameter-space objective is usually infeasible. Modern neural networks are high-dimensional, and different parameter settings can yield similar functions. More importantly, unlearned model auditing concerns whether unlearned samples still show membership traces in outputs or representations, instead of whether the parameters match a target distribution. Therefore, we move the auditing objective to the feature space. Instead of verifying the parameter distribution, we check whether $\mathcal{F}_u$ moves away from $\mathcal{F}_m$ and approaches $\mathcal{F}_n$. The feature-space auditing objective can then be written as the following Monge-type problem.

Definition 2.4. The Monge problem in feature space is:

$$\min_{T_F : T_{F\#}\mathcal{F}_u = \mathcal{F}_n} \int c_F\big(z, T_F(z)\big)\, d\mathcal{F}_u(z), \tag{5}$$

where T_F is the transport map in the auditing feature space, $T_{F\#}\mathcal{F}_u$ is the pushforward of $\mathcal{F}_u$ under T_F, and c_F measures the cost of transporting $\mathcal{F}_u$ to $\mathcal{F}_n$.

The Monge formulation does not require explicitly solving T_F. It only specifies the desired auditing direction: after sufficient unlearning, $\mathcal{F}_u$ should move away from $\mathcal{F}_m$ and approach $\mathcal{F}_n$. Thus, auditing only needs to measure how far $\mathcal{F}_u$ has moved toward the non-member state. As illustrated in Figure 2, we therefore relax the transport objective into a mixture-proportion problem and approximate $\mathcal{F}_u$ as a convex combination of $\mathcal{F}_m$ and $\mathcal{F}_n$. The relaxation of the conditions in this Monge problem is discussed in detail in Appendix B.1.

$$\mathcal{F}_u \approx (1-\rho)\mathcal{F}_m + \rho\mathcal{F}_n, \qquad \rho \in [0,1], \tag{6}$$

where ρ measures the non-member-like proportion in $\mathcal{F}_u$. Thus, the new unlearned model auditing objective is to estimate the mixture coefficient ρ instead of solving the feature-space transport map. In practice, ρ can be estimated by minimizing the discrepancy between empirical distributions.

Definition 2.5. The practical estimation problem for the mixture proportion is defined as:

$$\rho^* = \arg\min_{\rho\in[0,1]} D\left(\mathcal{F}_u, (1-\rho)\mathcal{F}_m + \rho\mathcal{F}_n\right), \tag{7}$$

where $D(\cdot,\cdot)$ denotes a distributional discrepancy measure. We then define the auditing objective as solving the optimization problem in Eq. (7), and use the resulting ρ^* as the auditing metric for the unlearned model.

3 Method

In this section, we propose Statistical Membership Inference (SMI) for reliable unlearned model auditing. Motivated by the previous analysis, SMI solves the optimization problem in Eq. (7). Statistical metrics for distributional distance are classical; see, e.g., Székely et al. [2007], Ramaswamy et al. [2016], Khamaru et al. [2025]. Appendices C and D provide the omitted mathematical details. We next introduce two solution strategies for SMI:

- **SMI:** An auditing tool that calculates low-order moments of distributions.
- **SMI-M:** An auditing tool that calculates the Maximum Mean Discrepancy (MMD) between distributions.

3.1 Statistical Membership Inference (SMI)

For the mixture-ratio estimation problem, SMI estimates the mixing proportion using the low-order statistics of the three distributions $\mathcal{F}_m$, $\mathcal{F}_n$, and $\mathcal{F}_u$. Under the assumption in Eq. (7), the mixed auditing distribution can be expressed as follows.

Proposition 3.1 (See Appendix C.3 for the proof.)**.** *If $\mathcal{F}_u$ is modeled as a mixture of $\mathcal{F}_n$ and $\mathcal{F}_m$ with proportions ρ and $1-\rho$, respectively, i.e.,*

$$\mathcal{F}_u = \rho\mathcal{F}_n + (1-\rho)\mathcal{F}_m, \tag{8}$$

then its mean μ_u and covariance Σ_u satisfy

$$\mu_u = \rho\mu_n + (1-\rho)\mu_m, \quad \Sigma_u = \rho\Sigma_n + (1-\rho)\Sigma_m + (\rho-\rho^2)(\mu_n-\mu_m)(\mu_n-\mu_m)^\top. \tag{9}$$

From Proposition 3.1, estimating the forgetting degree of $\mathcal{F}_u$ can be reduced to estimating the mixture coefficient ρ. SMI estimates ρ by matching the empirical covariance of $\mathcal{F}_u$ with the covariance predicted by the mixture model.

Lemma 3.2 (Solution of SMI)**.** *Let $d = \mu_n - \mu_m$, $B = dd^\top$, $A = \Sigma_u - \Sigma_m$, $G = \Sigma_n - \Sigma_m + B$. SMI estimates the mixture ratio by solving:*

$$\min_{0\le\rho\le1} \left\|\Sigma_u - \left[\rho\Sigma_n + (1-\rho)\Sigma_m + (\rho-\rho^2)(\mu_n-\mu_m)(\mu_n-\mu_m)^\top\right]\right\|_F^2. \tag{10}$$

The solution is selected from the search space $\mathcal{C}$:

$$\rho^\star = \arg\min_{\rho\in\mathcal{C}} \left\|A - \rho G + \rho^2 B\right\|_F^2. \tag{11}$$

The search space $\mathcal{C}$ is:

$$\mathcal{C} = \{0,1\} \cup \Big\{ \rho \in [0,1] : 2\|B\|_F^2 \rho^3 - 3\langle G, B\rangle_F \rho^2 + \big(\|G\|_F^2 + 2\langle A, B\rangle_F\big)\rho - \langle A, G\rangle_F = 0 \Big\}, \quad (12)$$

where $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$ denotes the Frobenius inner product. Specifically, the solution process can be described as first solving a univariate cubic equation in ρ, and then substituting the obtained roots into the search space to select the minimizer. The complete proof is provided in Appendix C.4.

The SMI optimization is simple, and the resulting ρ^* serves as the final auditing result. SMI computes the mean and variance of neural features, rather than raw data. For a sample x and model parameters w, we denote the extracted feature by $g(x; w)$. Algorithm 1 gives the pseudo-code of SMI. The optimization in Lemma 3.2 may be unstable because it depends on the empirical statistics Σ_u, Σ_n, and Σ_m. With limited auditing samples, covariance estimates can fluctuate, especially for high-dimensional neural features. Direct covariance matching may therefore give unstable mixture-ratio estimates. To improve stability, we instead use kernel mean embeddings and match the embedding of $\mathcal{F}_u$ to a convex combination of those of $\mathcal{F}_m$ and $\mathcal{F}_n$.

Algorithm 1: SMI

Input: Member data D_m, non-member data D_n, pending-audit data D_u, model w, audit function $g(\cdot, w)$

Output: Mixture ratio ρ^*.

$(\mu_m, \Sigma_m) \leftarrow \mathrm{stats}(g(D_m; w))$;

$(\mu_n, \Sigma_n) \leftarrow \mathrm{stats}(g(D_n; w))$;

$(\Sigma_u) \leftarrow \mathrm{stats}(g(D_u; w))$;

$\rho^\star \leftarrow \arg\min_{\rho \in \mathcal{C}} \left\| A - \rho G + \rho^2 B \right\|_F^2$;

return ρ^*;

3.2 Statistical Membership Inference-MMD (SMI-M)

The computational error in estimating the mixture ratio using low-order moments mainly arises from the estimation of the covariance matrix. Therefore, we introduce a statistical feature optimization method, embedding $g(\cdot; w)$ into a RKHS, thereby transforming the original second-order moment (covariance Σ) optimization problem into a first-order moment embedding problem in RKHS.

Specifically, we introduce a kernel function $k(\cdot, \cdot)$ for $g(\cdot; w)$ and define the kernel mean embedding as $\mu_\cdot^{(k)} = \mathbb{E}[k(g(\cdot; w), g(\cdot; w)^\top)]$. This mapping transforms the second-order information of the original distribution into the first-order information of the embedded distribution. Accordingly, the Lemma 3.2 can be reformulated as follows:

Lemma 3.3 (Optimization problem of SMI-M)**.** *For the embedded $\mu_f^{(k)}, \mu_v^{(k)}, \mu_t^{(k)}$, the optimization problem (10) is equivalent to the following convex quadratic programming problem:*

$$\min_{0 \le \rho \le 1} \ \rho^2 \|\mu_v^{(k)} - \mu_t^{(k)}\|_{\mathcal{H}}^2 - 2\rho \langle \mu_f^{(k)} - \mu_t^{(k)}, \mu_v^{(k)} - \mu_t^{(k)} \rangle_{\mathcal{H}}, \quad (13)$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the Hilbert space norm, and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the Hilbert space inner product. This problem is a convex quadratic programming problem and can be solved directly by a solver without manually deriving a closed-form solution. The complete proof is provided in Appendix C.5.

3.3 Resampling-based Uncertainty Estimation for SMI

Inspired by shadow models in MIA, we use resampling to estimate the uncertainty of SMI. Shadow-model methods train multiple auxiliary models, which is costly for unlearned model auditing. SMI avoids this cost by bootstrapping D_u in the auditing feature space. For each resampled set, we solve the SMI optimization again and obtain $\{\rho_b^*\}_{b=1}^B$.

Resampling doesn't change the point estimate ρ^*. It only provides an empirical reference range for finite-sample sensitivity. Concentrated ρ_b^* indicates stable auditing, while large variation suggests that ρ^* should be interpreted with caution. We report $\{\rho_{5\%}^*, \rho_{95\%}^*\}$ as the reference range.

4 Experiment

To verify the effectiveness of SMI, we will examine it from the following five perspectives:

1. **Performance:** *Does SMI outperform conventional MIA-based auditing?*In Section 4.2, we evaluate SMI on public datasets and the unlearning-audit benchmark, and compare it with conventional MIA methods as baselines.
2. **Computational Cost:** *Does SMI avoid the substantial computational overhead of shadow-model-based auditing?* In Section 4.3, we compare the runtime and inference cost of different SMI variants and conventional MIA methods, with a particular focus on shadow model and non-shadow model settings.
3. **Limited-sample Reliability:** *Can SMI provide stable auditing results with limited unlearning samples?* In Section 4.4, we evaluate whether SMI remains reliable when the number of audited samples is limited, and analyze the stability of its estimated forgetting ratio and resampling-based reference range.
4. **Robustness:** *Is SMI-M stable under different experimental conditions?* In Section 4.5, we conduct robustness experiments for SMI-M to assess whether its auditing performance remains stable under variations in data, model behavior, or sampling conditions.
5. **Practicality:** *Does SMI work in real unlearning scenarios?* In Section 4.6, we evaluate SMI in realistic machine unlearning settings to examine whether it can reliably audit actual forgetting behavior beyond controlled benchmark comparisons.

4.1 Experimental Setup

Dataset and Models: We construct image-classification tasks using the CIFAR-10, CIFAR-100 Krizhevsky et al. [2009], and CINIC-10 Darlow et al. [2018] datasets. We adopt ResNet-18 He et al. [2016], Resnet50 Koonce [2021] and VIT Yuan et al. [2021] for both the target model and the shadow models required by conventional MIA baselines.

Baselines: We evaluate model-agnostic MIA methods using the unlearning audit benchmark released alongside IAM Wang et al. [2025a]. We adopt seven MIA-based strategies as baselines forunlearned model auditing, as listed below:

- **Random:** Randomly estimates the member proportion as a worst-case reference.
- **Unleak (2021) :**This metric measures the output discrepancy before and after unlearning.
- **LiRA (2022):** Performs auditing by comparing the output likelihood ratio between the unlearning dataset and the training dataset on the target model.
- **EMIA (2022):** Extends LiRA with incremental auditing decisions.
- **RMIA (2024):** Builds a robust likelihood ratio based on LiRA.
- **IAM (2025a):** Uses a model interpolation strategy to localize model information for auditing.
- **RULI (2025):** Explicitly models the unlearning process for auditing.

MIA Settings. For all MIA baselines, we use the online white-box setting, where the attacker accesses the pre-unlearning model, the post-unlearning model, and all shadow models. This gives the baselines the strongest capability under our auditing setup. For LiRA-type methods, we use the full logit vector and learn the threshold on an independent calibration set with 10,000 member and 10,000 non-member samples. The learned threshold is then applied to the unlearning data, with calibration and evaluation strictly separated. For MIA methods with class-level decisions, we report the better result between class-level and full-sample evaluation as a strong upper bound.

Metrics. MIA methods output sample-level membership decisions, while SMI methods estimate a distribution-level mixture ratio. For MIA, we report FNR, FPR, and the proportion of unlearning samples classified as non-members as the aligned ρ^*. For SMI, we report the estimated ρ^*.

4.2 The Performance of SMI

Unlearning Task Settings. In this subsection, we use full retraining as the optimal unlearned model auditing target, and set the unlearning data to a randomly selected 5% or 10% subset of each dataset. Practical unlearning algorithms are discussed in the following subsection.

Table 1: Auditing performance of SMI and MIA-based methods on complete unlearning. The Dataset column lists the model, task, unlearning percentage, and number of shadow models. Experimental results for some datasets, with the **best** and second best results are highlighted. The complete results can be found in Appendix E.

Dataset	Metric	SMI	SMI-M	RULI	IAM	RMIA	EMIA	LiRA	Unleak	Ramdom
Resnet18-Cifar10-5%-64	FNR	–	–	13.28%	12.72%	49.68%	70.82%	13.80%	41.24%	49.52%
	FPR	–	–	0.00%	13.08%	50.04%	70.72%	0.00%	43.48%	50.88%
	ρ^*	88.89%	**92.44%**	87.94%	88.67%	66.68%	39.10%	87.84%	62.58%	50.06%
Resnet18-Cifar100-5%-16	FNR	–	–	15.60%	17.76%	87.80%	24.72%	10.47%	81.04%	48.72%
	FPR	–	–	0.00%	17.96%	87.96%	22.24%	0.00%	82.88%	49.52%
	ρ^*	79.64%	**89.84%**	83.00%	89.36%	67.19%	84.64%	86.64%	28.52%	50.88%
Resnet50-Cifar100-5%-32	FNR	–	–	36.76%	86.72%	91.92%	66.52%	37.41%	93.08%	47.68%
	FPR	–	–	0.00%	63.83%	56.00%	57.56%	0.00%	47.69%	50.15%
	ρ^*	81.22%	**92.64%**	73.08%	63.56%	58.24%	67.84%	87.04%	10.40%	54.68%
Resnet50-Cifar100-10%-32	FNR	–	–	26.90%	87.73%	91.50%	70.72%	38.29%	93.56%	49.71%
	FPR	–	–	0.00%	61.69%	53.21%	57.81%	0.00%	49.77%	42.27%
	ρ^*	83.46%	**93.95%**	74.44%	64.98%	60.74%	69.94%	87.38%	4.68%	54.14%
Resnet50-Cinic10-10%-16	FNR	–	–	11.36%	29.92%	11.88%	22.08%	11.18%	19.46%	49.82%
	FPR	–	–	0.00%	30.96%	11.28%	21.68%	0.00%	14.04%	50.23%
	ρ^*	90.89%	**92.27%**	89.88%	85.35%	87.64%	89.66%	90.47%	78.02%	49.76%
VIT-Cinic10-10%-16	FNR	–	–	15.92%	57.40%	85.04%	65.36%	18.76%	43.24%	49.97%
	FPR	–	–	0.00%	53.37%	86.12%	65.24%	0.00%	44.76%	51.93%
	ρ^*	91.44%	**93.62%**	89.84%	41.01%	15.33%	36.66%	83.51%	55.71%	51.04%

Results. Table 1 reports the results across datasets for both MIA- and SMI-based auditing. Overall, SMI-M achieves the best estimation accuracy. Among non-SMI methods, RuLI and LiRA perform best in most settings: RuLI explicitly models the unlearning data, while LiRA uses a simpler and more stable auditing pipeline.Better auditing performance is usually associated with lower FPR; for example, RuLI and LiRA consistently achieve an FPR of 0. However, different metrics are not strictly aligned. In many settings, LiRA has a higher FNR than RuLI, whereas RuLI gives a more reliable ρ^*. This supports our corollary: even a well-trained MIA may still suffer from an additional auditing error in unlearned model auditing.

4.3 The Computational Cost of SMI

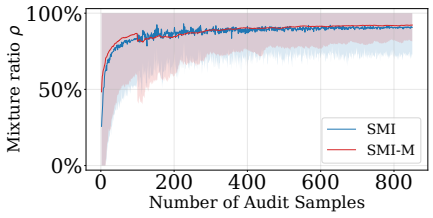
Computational Cost. Table 2 reports the computational overhead of different MIA methods and SMI variants for a single auditing run on CIFAR-100. For SMI methods, the runtime covers auditing over all classes. Methods without shadow models generally have low cost, although SMI-M is slower due to its kernel embedding computation.This overhead is still small compared with training shadow models. For example, training one LiRA shadow model with a ResNet-18 backbone takes enough time to run SMI-M hundreds of times, while shadow-model-based MIA usually requires multiple shadow models. For SMI, the main cost comes from bootstrapping, and the runtimes in Table 2 are reported with 200 bootstrap iterations. Non-shadow methods require one inference pass, except Unleak, which requires two. Shadow-model methods require one pass for each shadow model plus one additional pass.

Table 2: Computational cost of MIA methods. A dash (–) indicates that the corresponding method is not used. The compute time reported for Train shadow model refers to the cost of training a single shadow model.

Method	Compute times	Shadow model	Bootstrap
Inference on one model	4.47×10^4ms	-	-
Logit Metric	217.25ms	-	-
Unleak	667.09ms	-	-
SMI	930.71ms	-	200
SMI-M	4301.66ms	-	200
Train one shadow model	2.46×10^6ms	-	-
LiRA	1464.51ms	8	-
EMIA	1625.26ms	8	-
RMIA	2475.42ms	8	-
IAM	2944.35ms	8	-
RULI	5672.83ms	8	-

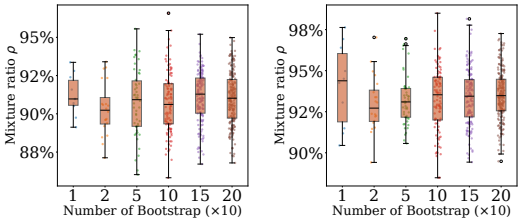
4.4 The Small-Sample Reliability of SMI

The Small-Sample Challenge of SMI. Since SMI relies on distributional mixture estimation, its error increases with smaller sample sizes. We therefore study its sample requirement on CINIC-10 by varying the audit size from 1 to 800. For each size, we run 1,000 resampling audits when the size



(a) CINIC-ResNet50

Figure 3: Effect of audit sample size on SMI and SMI-M performance. Solid lines show the mean, and dashed lines show the reference range.



(a) SMI (b) SMI-M

Figure 4: Performance fluctuation of SMI and SMI-M with different numbers of resampling iterations on 10% randomly unlearned data.

Table 3: Relationship between different kernels and the performance of SMI-M. Values in parentheses denote the multiplier applied to the median bandwidth.

Data	RBF	RBF(0.25X)	RBF(0.5X)	RBF(2X)	RBF(4X)	Laplacian	Linear	Polynomial	Consine
Cifar100-5%	89.84%	24.39%	77.70%	86.88%	86.72%	80.05%	86.65%	82.56%	83.98%
Cifar100-10%	91.84%	26.03%	79.22%	89.88%	89.90%	82.41%	89.88%	85.63%	88.16%

is below 200 and 500 audits otherwise. As shown in Figure 3, SMI converges around 200 samples, or about 20 samples per class for a 10-class dataset.

Bootstrap Overhead. Figure 4 illustrates the auditing performance of SMI and SMI-M under different numbers of bootstrap groups. As a white-box auditing method, SMI has already demonstrated strong and reliable performance advantages. Across varying numbers of bootstrap groups, SMI does not exhibit significant fluctuations in either the median performance or the reference ranges, indicating that the bootstrap procedure is well controlled and does not need to be excessively increased to improve SMI's performance.

4.5 The Robustness of SMI

Hyperparameter. Since SMI-M adopts a kernel embedding strategy, it involves several hyperparameters that need to be discussed. The hyperparameters of MMD generally include the choice of kernel function and the bandwidth selection strategy. In this subsection, we consider the RBF kernel, Laplacian kernel, linear kernel, polynomial kernel, and cosine kernel, as well as different RBF-based bandwidth selection strategies. The experimental results are shown in the table 3. Overall, the standard median bandwidth with the RBF kernel achieves the best MMD hyperparameter setting. The bandwidth is computed as the median distance over 10,000 sampled points from the member, non-member, and unlearning data. If the number of unlearning samples is small, the bandwidth can be estimated using only the member and non-member data.

Audit features: Since SMI depends strongly on feature choice, we supplement the main experiments by comparing logits, logit gradients, their concatenation, and loss. The experimental results are shown in the table 4. We adopt logit gradients for their best overall performance.

4.6 The Practicality of SMI

Practical Tasks. Beyond retraining, practical scenarios involve more approximate unlearning tasks that are difficult to audit. These methods may exhibit low accuracy on D_u, but their true level of forgetting is hard to determine. In this subsection, we evaluate the unlearning performance of three approximate unlearning strategies other than retraining: FineTune, Fisher, and Scrub. Their detailed descriptions are provided in Appendix E.3. The final results are shown in Table 5. Among them, Scrub achieves the best unlearning performance.

Specialized Auditing Tasks. Since our main experiments focus on in-distribution auditing, we further evaluate SMI in four out-of-distribution (OOD) settings: PathMNIST samples audited with a model trained on CIFAR-10; backdoor samples audited with a clean CIFAR-10 model; specific

Table 4: The auditing scores of SMI and SMI-M are compared across different features at a five percent unlearning ratio. Feature dimensions in parentheses correspond to CIFAR-10 and CIFAR-100, respectively.

Audit feature	CIFAR-10		CIFAR-100	
	SMI	SMI-M	SMI	SMI-M
Logit gradients (10, 100)	88.89%	92.44%	81.22%	92.64%
Logits (10, 100)	82.41%	91.53%	74.17%	88.79%
Logit gradients + logits (20, 200)	84.17%	94.68%	71.26%	92.83%
Loss (1, 1)	64.73%	–	59.83%	–

Table 5: SMI-based Detection of Unlearning Levels in Practical Unlearning Tasks

Method	Cifar100				Cinic			
	Resnet18-5%	Resnet18-10%	Resnet50-5%	Resnet50-10%	Resnet50-5%	Resnet50-10%	Vit-5%	Vit-10%
Retrain	89.84%	91.84%	92.64%	93.95%	90.72%	92.27%	92.13%	93.62%
FineTune	63.01%	66.54%	59.63%	61.79%	67.35%	68.66%	58.99%	55.42%
Fisher	32.51%	34.70%	39.82%	34.86%	40.83%	44.98%	44.77%	38.93%
Scrub	76.34%	75.95%	72.61%	77.58%	69.21%	63.27%	62.43%	66.16%

corpora audited with a language model trained on different languages; and adversarially generated samples. In each setting, we test whether the target data participated in training. As shown in Table 6, SMI remains effective in most tested OOD settings, except for adversarial samples.

Table 6: The table reports auditing estimates and fitting objective values for SMI and SMI-M under distribution shifts and adversarial perturbations. Estimates are expressed as percentages in the first three rows and as raw values in the final row, which presents a failure case.

Scenario	SMI		SMI-M	
	ρ^*	Objective	ρ^*	Objective
Cross-domain images	74.63%	0.87	81.44%	0.13
Backdoor samples	87.10%	0.15	93.48%	0.11
Cross-language corpora	68.59%	0.77	79.63%	0.64
Gradient-based adversarial samples	13.50	64.3	64.82	127.5

5 Conclusion

In this work, we propose SMI, an unlearned model auditing method based on measuring distributional distances. We revisit the implicit bounds underlying MIA attacks and MIA-based auditing, and argue that failed attacks can produce illusory forgetting. To address this issue, we propose a mixture-ratio estimation strategy for auditing model unlearning, which provides an efficient unlearned model auditing approach without requiring shadow models. Experimental results show that our MMD-based SMI-M outperforms existing MIA-based methods. It demonstrates strong capability in reliably measuring unlearning outcomes, achieving efficient computation without training shadow models, and providing reference intervals to assess the reliability of the auditing conclusions.

Acknowledgements

We thank Xinpeng Ling, Zhanpeng Shi, Wei Mo, Sizhe Liu, Baolin Chen, Yihao Xiao, and Kejia Zhang for helpful and fruitful discussions. This work is funded by the National Natural Science Foundation of China (Grant No. 62203309).

References

Li Bai, Qingqing Ye, Xinwei Zhang, Sen Zhang, Zi Liang, Jianliang Xu, and Haibo Hu. Toward efficient inference attacks: Shadow model sharing via mixture-of-experts. *arXiv preprint arXiv:2510.13451*, 2025. 2

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022. 7, 14

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 896–911, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384544. doi: 10.1145/3460120.3484756. URL https://doi.org/10.1145/3460120.3484756. 3, 7, 14

Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018. 7

Yuntao Du, Jiacheng Li, Yuetian Chen, Kaiyuan Zhang, Zhizhen Yuan, Hanshen Xiao, Bruno Ribeiro, and Ninghui Li. Cascading and proxy membership inference attacks. *arXiv preprint arXiv:2507.21412*, 2025. 2

Ali Ebrahimpour-Boroojeny, Hari Sundaram, and Varun Chandrasekaran. Not all wrong is bad: Using adversarial examples for unlearning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 14950–14971. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ebrahimpour-boroojeny25a.html. 2, 14

Jie Fu, Yuan Hong, Zhili Chen, and Wendy Hui Wang. Safeguarding graph neural networks against topology inference attacks. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 2144–2158, 2025. 1

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *Ieee Access*, 7:47230–47244, 2019. 2

Ling Han, Hao Huang, Dustin Scheinost, Mary-Anne Hartley, and María Rodríguez Martínez. Unlearning information bottleneck: Machine unlearning of systematic patterns and biases. *arXiv preprint arXiv:2405.14020*, 2024. 2

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 497–519. IEEE, 2025. 1

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

Jean Honorio and Tommi Jaakkola. Tight bounds for the expected risk of linear classifiers and pac-bayes finite-sample guarantees. In *Artificial Intelligence and Statistics*, pages 384–392. PMLR, 2014. 3

Koulik Khamaru, Yash Deshpande, Tor Lattimore, Lester Mackey, and Martin J Wainwright. Near-optimal inference in adaptive linear regression. *The Annals of Statistics*, 53(6):2329–2355, 2025. 4, 5

Brett Koonce. Resnet 50. In *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72. Springer, 2021. 7

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009. 7

Uyen N Le-Khac and Vinh NX Truong. A survey on large language models unlearning: taxonomy, evaluations, and future directions. *Artificial Intelligence Review*, 58(12):399, 2025. 1, 14

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14, 2025. 2

Yang Liu, Mingyuan Fan, Cen Chen, Ximeng Liu, Zhuo Ma, Li Wang, and Jianfeng Ma. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pages 280–289. IEEE, 2022a. 2

Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pages 1749–1758. IEEE, 2022b. 1, 14

Yi Liu, Weixiang Han, Chengjun Cai, Xingliang Yuan, and Cong Wang. Privtune: Efficient and privacy-preserving fine-tuning of large language models via device-cloud collaboration, 2026. URL https://arxiv.org/abs/2512.08809. 1

Zihao Luo, Xilie Xu, Feng Liu, Yun Sing Koh, Di Wang, and Jingfeng Zhang. Privacy-preserving low-rank adaptation against membership inference attacks for latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5883–5891, 2025. 3

Ryoto Miyamoto, Xin Fan, Fuyuko Kido, Tsuneo Matsumoto, and Hayato Yamana. Openlvlm-mia: A controlled benchmark revealing the limits of membership inference attacks on large vision-language models. *arXiv preprint arXiv:2510.16295*, 2025. 2

Nima Naderloui, Shenao Yan, Binghui Wang, Jie Fu, Wendy Hui Wang, Weiran Liu, and Yuan Hong. Rectifying privacy and efficacy measurements in machine unlearning: A new inference attack perspective. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 5545–5564, 2025. 2, 7, 14

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753, 2019. doi: 10.1109/SP.2019.00065. 2

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025. 1, 14

Serena Nicolazzo, Antonino Nocera, et al. How secure is forgetting? linking machine unlearning to machine learning attacks. *arXiv preprint arXiv:2503.20257*, 2025. 2

Yuki Ohnishi and Jean Honorio. Novel change of measure inequalities with applications to pac-bayesian bounds and monte carlo estimation. In *International conference on artificial intelligence and statistics*, pages 1711–1719. PMLR, 2021. 3

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International conference on machine learning*, pages 2052–2060. PMLR, 2016. 5

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. 2

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. 2007. 5

Cheng-Long Wang, Qi Li, Zihang Xiang, Yinzhi Cao, and Di Wang. Towards lifecycle unlearning commitment management: Measuring sample-level unlearning completeness. *arXiv preprint arXiv:2506.06112*, 2025a. 2, 7, 14

Zhiqi Wang, Chengyu Zhang, Yuetian Chen, Nathalie Baracaldo, Swanand R Kadhe, and Lei Yu. Membership inference attacks as privacy tools: Reliability, disparity and ensemble. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 1724–1738, 2025b. 2

Rongzhe Wei, Mufei Li, Mohsen Ghassemi, Eleonora Kreačić, Yifan Li, Xiang Yue, Bo Li, Vamsi K. Potluru, Pan Li, and Eli Chien. Underestimated privacy risks for minority populations in large language model unlearning, 2025. URL https://arxiv.org/abs/2412.08559. 2

Chengcan Wu, Zeming Wei, Huanran Chen, Yinpeng Dong, and Meng Sun. Reliable unlearning harmful information in llms with metamorphosis representation projection. In *NeurIPS Workshop on Reliable ML from Unreliable Data*, 2025. 2

Dayong Ye, Tianqing Zhu, Jiayang Li, Kun Gao, Bo Liu, Leo Yu Zhang, Wanlei Zhou, and Yang Zhang. Data duplication: A novel multi-purpose attack paradigm in machine unlearning, 2025. URL https://arxiv.org/abs/2501.16663. 1

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pages 3093–3106, 2022. 7

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 7

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58244–58282. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zarifzadeh24a.html. 2, 7, 14

Chenhao Zhang, Muxing Li, Feng Liu, Weitong Chen, and Miao Xu. Unlearning evaluation through subset statistical independence. *arXiv preprint arXiv:2603.00587*, 2026. 14

Qingjie Zhang, Haoting Qian, Zhicong Huang, Cheng Hong, Minlie Huang, Ke Xu, Chao Zhang, and Han Qiu. Understanding the dilemma of unlearning for large language models. *arXiv preprint arXiv:2509.24675*, 2025. 2

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. *arXiv preprint arXiv:2501.13453*, 2025. 2

Overview

For a detailed exposition of SMI, this appendix is organized into five parts: Related Work, Disscussion and Limitations, Theory Proofs, Theoretical Elaboration on SMI-M and MMD and The Complete Experimental.

A Related Work

A.1 Machine unlearning

Machine unlearning aims to selectively remove the influence of the data to be unlearned, D_u, from a trained model. Let the initial model w_0 be trained on the member training data D_m, where D_u denotes the unlearning set to be removed and D_n denotes non-member data that have never participated in training Ebrahimpour-Boroojeny et al. [2025]. The goal of machine unlearning is to eliminate the influence of D_u while preserving the model utility on D_m as much as possible, so that the unlearned model w_u becomes close to the retrained model w_r obtained by training only on $D_m \setminus D_u$ Le-Khac and Truong [2025], Nguyen et al. [2025], Liu et al. [2022b].

A.2 Auditing Strategies for Machine Unlearning

A common privacy-oriented evaluation strategy for machine unlearning is to regard a model trained from scratch without the unlearning data as the optimal unlearned model. This view is based on the premise that, if an unlearning algorithm truly removes the influence of the unlearning data, the resulting model should behave similarly to this retrained model. Following this idea, existing studies usually evaluate unlearning by comparing the unlearned model and the retrained model in terms of parameters, output distributions, prediction confidence, or task performance Zhang et al. [2026], Nguyen et al. [2025].

Another intuitive privacy-oriented evaluation method is MIA. In the context of machine unlearning, Chen et al. first showed that unlearning itself may introduce new privacy risks: by comparing the outputs of the original model and the unlearned model, an attacker can infer whether a target sample was included in the original training set but removed from the unlearned model Chen et al. [2021]. Beyond this unlearning-specific leakage, Carlini et al. formulated membership inference from first principles and proposed a likelihood-ratio attack that provides a stronger and more systematic way to detect membership traces Carlini et al. [2022]. Zarifzadeh et al. further improved the practicality of MIA by developing a low-cost, high-power attack that remains effective even with limited reference models Zarifzadeh et al. [2024]. More recently, Wang et al. argued that binary membership inference alone is insufficient for measuring approximate unlearning, and proposed sample-level unlearning completeness to characterize how completely each individual sample has been forgotten Wang et al. [2025a]. Naderloui et al. further revisited privacy and efficacy measurements in machine unlearning from a new inference attack perspective, showing that unlearning evaluation requires attacks tailored to the post-unlearning setting rather than relying only on coarse aggregate metrics Naderloui et al. [2025]. Together, these studies motivate MIA-based evaluation as an interpretable privacy audit: if the unlearning data have been effectively removed, they should no longer exhibit membership traces that allow an attacker to distinguish them from non-member samples.

B Discussion and Limitations

To further elucidate the workings of SMI, we address potential queries regarding SMI in advance to facilitate a better understanding of its underlying philosophy:

Question 1: The contribution of SMI appears to be merely replacing the binary classifier with a statistical metric discriminator; this contribution seems incremental.

Answer: From a technical standpoint, the contribution of SMI differs significantly from traditional MIA methods, which focus on constructing a $g(x; w)$ with superior performance. The core contribution of SMI is the proposal of a novel perspective: When the attack on a single sample becomes unreliable as an evaluation metric, directly measuring statistical indicators for the entire audit population is a more reliable approach. In reality, indistinguishable samples always exist; however,

when viewed through the lens of statistical measurement, they are naturally categorized within the distribution of member samples. Conversely, if these indistinguishable samples are involved in the training of a binary classifier as non-member data, they disrupt the classifier's training process.

Question 2: Why not consider using Wasserstein methods to improve mixture proportion estimation, given their natural alignment with the Monge problem?

Answer: This is an incisive observation. We initially adopted SMI-W as our working strategy, hoping to formulate it as an improvement over SMI-M. However, our experiments showed that its performance was inferior to the simple estimation strategy used by SMI.

The formalization of the Monge problem helped us understand why the Wasserstein approach failed in this setting. Wasserstein distance measures the optimal discrepancy between two distributions. In our context, it can be viewed as a metric for an optimal forgetting strategy defined over the parameter space, or equivalently, as measuring an optimal transformation process in a modified parameter or feature space. However, neither retraining nor fine-tuning a model fundamentally satisfies the geometric assumptions underlying the Wasserstein formulation: the actual update trajectory does not align with the optimal transport geodesic.

Put more simply, Wasserstein distance assumes that there exists a shortest geodesic between two distributions and measures the length of that path. Yet the forgetting behavior we truly need to quantify does not lie on this path; from the perspective of the Wasserstein metric, all feasible forgetting strategies we can conceive of are too crude. This also motivates our future work: in high-dimensional spaces, Wasserstein metrics and gradient flows may not necessarily be aligned, and such statistical distances must therefore be used with caution.

Question 3: What advantages does SMI offer over MIA as an auditing strategy in practical audit settings?

Answer: We believe the most important advantage is that, during the auditing process, one does not need to worry about the effectiveness of the auditing tool itself. According to our experimental results, the greatest strength of SMI-M is its exceptional stability: it does not exhibit large performance fluctuations during auditing.

By contrast, a major limitation of MIA-based methods is that, before conducting the audit, one must assess whether the designed MIA discriminator has been sufficiently trained. In practice, however, it is difficult to determine whether MIA training is truly sufficient. Our experiments further show that, under the same FPR, a lower FNR does not necessarily lead to more accurate MIA decisions.

The SMI family provides a set of accurate and efficient auditing metrics whose error scales almost solely with the size of the unlearning set. Before applying SMI to a real forgetting audit, one can directly evaluate its task-specific performance on a test set.

Question 4: I noticed that the PAC-Bayes theory you use appears to differ from the commonly used KL-divergence-based formulation. Could you elaborate on this difference?

Answer: We recognize that this issue is of particular interest to researchers working on the theoretical side, and we would like to clarify why we adopt the χ^2 divergence.

At the beginning, we attempted to use the most classical PAC-Bayes theorem based on KL divergence. However, while developing the proof, our literature review led us to a PAC-Bayes strategy based on χ^2 divergence. We found that this formulation is more useful for decomposing the statistical error and deriving the resulting auditing error.

More interestingly, it is also better aligned with the methodology underlying our experiments. In essence, our approach solves the distributional mixture problem by exploiting higher-order statistical moments. This inevitably requires a statistical measure that can capture higher-order moment information. In this setting, χ^2 divergence is substantially more effective than KL divergence at capturing second-order moment behavior.

Limitation: As the saying goes, there is no free lunch. Since SMI does not use shadow models as auxiliary information, it must inevitably cope with the noise introduced by small sample sizes. As a statistical strategy, SMI relies on having a sufficient number of samples. Its auditing performance on a single sample is weaker than that of MIA methods, which can exploit both the sample itself and

information from shadow models. Empirically, our results suggest that at least 20 samples are needed to conduct an effective audit for a given class.

B.1 Discussion on the Relaxation of the Monge Problem

We note that some readers may have questions about the relaxed formulation in Eq. (14) derived from the Monge problem, and may question whether it is valid. We clarify this issue in this section.

$$\mathcal{F}_u \approx (1-\rho)\mathcal{F}_m + \rho\mathcal{F}_n, \qquad \rho \in [0,1]. \tag{14}$$

Specifically, $\mathcal{F}_u$, $\mathcal{F}_m$, and $\mathcal{F}_n$ denote the feature distributions of the audited forgotten samples, member reference samples, and non-member reference samples, respectively. Ideally, a transport map $T_{\mathcal{F}}$ pushes $\mathcal{F}_u$ to $\mathcal{F}_n$, yielding a Monge problem in the auditing feature space. The corresponding optimal transport path describes an ideal process in which each forgotten sample's features move toward the corresponding non-member features.

Actual unlearning, however, is driven by model parameter updates. Along a parameter path $\{\theta_t\}$, feature changes are determined by $g(\cdot;\theta_t)$. This process generally neither minimizes the feature-space transport cost nor ensures that samples follow the directions defined by the optimal transport map. Thus, although $W_2(\mathcal{F}_u, \mathcal{F}_n)$ measures distributional discrepancy, it cannot be directly interpreted as the forgetting ratio or progress induced by parameter updates.

To account for this geometric mismatch, we characterize distributional changes through statistical matching. The squared Wasserstein distance admits the following moment-based lower bound:

$$W_2^2(P,Q) \geq \|\mu_P - \mu_Q\|_2^2 + d^2(\Sigma_P, \Sigma_Q), \tag{15}$$

where μ_P, μ_Q and Σ_P, Σ_Q denote the means and covariances of the two distributions, and $d^2(\Sigma_P, \Sigma_Q)$ is the covariance distance term. This bound motivates comparing distributions through their means and covariances, but does not directly imply the convex-mixture relation in Eq. (14).

Equation (14) is a heuristic modeling approximation motivated by this perspective. It represents the audited distribution as a mixture of member and non-member reference distributions, replacing transport-map estimation with mixture-proportion estimation. Under this approximation, ρ quantifies the non-member-like component of the audited distribution. Its suitability for a particular unlearning scenario is assessed empirically.

First, this condition does not always hold. We construct a counterexample: when a malicious attacker uses adversarial example generation to forge the unlearning process, we find that the estimated $\rho^* \gg 10$. This demonstrates that SMI-based unlearned model auditing can quickly fail under adversarial perturbations. Nevertheless, the abnormal value of ρ^* can still indicate that an anomalous result has occurred in the current unlearning process.

In addition, the objective loss obtained when solving the optimization problem can serve as a classical evaluation metric. After normalizing the input features, we find that the objective value is on the order of 10^{-1}, which indirectly supports that the mixture-proportion assumption is likely to hold in practical scenarios.

Finally, we would like to provide an insight from the perspective of the statistical estimation process. In high-dimensional statistical metrics, we often observe that part of the computation of the Wasserstein distance can be related to mixture-ratio estimation. From this viewpoint, SMI can be regarded as a Wasserstein-like high-dimensional statistical measurement strategy. Moreover, for neural network output features with relatively low signal-to-noise ratios, capturing higher-order moments may not be particularly effective. Therefore, we are confident in this heuristic conclusion.

Here, we must clarify that if one aims to obtain a strict convexity result, additional complex assumptions would be required for the proof. However, such a proof would provide limited practical guidance, and its significance would be much weaker than our decomposition of the auditing error. In future work, we will continue along the direction of AI interpretability and attempt to analyze the behavior of neural networks in feature space, although such conclusions are not easy to obtain.

C Theory Proofs

Regarding the detailed exposition of SMI, this appendix is organized into three sections: related discussions on the SMI framework, theoretical proofs presented in the SMI paper, and further elaboration on the MMD distance.

We extend our gratitude to the readers for reaching this point. The core philosophy of SMI stems from PAC-Bayes theory, long-tailed distribution theory, and unsupervised learning theory. In the proof and characterization of our theorems, we adopt many ideas from these works. First and foremost, we thank every reader interested in SMI theory. We look forward to building the Unlearning Audit community with you.

Specifically, the empirical risk of MIA attacks is inspired by PAC theory and MIA prior theory, leading us to construct a PAC error form for MIA. As shown in Theorem C.1, this constitutes one of the core contributions of our paper. Considering that this contribution is difficult to describe simply in the main text, we discuss in detail in this appendix its impact on our work and its guiding significance for MIA auditing tools.

Learning theory investigates the gap between training error and generalization error, providing performance guarantees for a well-trained learner on new data. A model with strong generalization capability will have a theoretical generalization error that is as small as possible.

Consider a learning algorithm setting where there exists a sample set $S = \{z_1, z_2, \cdots, z_m\}$ containing m samples. All these samples are drawn from the same **unknown distribution** $\mathcal{D}$, which is a probability measure on a measurable space $\mathcal{Z}$. For a supervised learning problem, this measurable space can be decomposed into a feature part and a label part, i.e., $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the label space. If the problem is a binary classification problem, then $\mathcal{Y} = \{0, 1\}$; if it is a regression problem, then $\mathcal{Y} = \mathbb{R}$. Given a measurable hypothesis space $\mathcal{H}$ and a loss function $l : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, the general learning objective is to find the best hypothesis $h \in \mathcal{H}$ that minimizes the **true risk** (also known as the **expected risk**):

$$R_D(h) = \mathbb{E}_{z\sim\mathcal{D}} l(h, z) \tag{16}$$

For classification problems, $\mathcal{H}$ can be a class of classifiers. If $\mathcal{H}$ is a set of parameterizable classifier models, then h can be a weight vector. Let $\mathcal{M}(\mathcal{H})$ denote a **probability measure space** over $\mathcal{H}$. To incorporate model uncertainty, we consider using **stochastic inference** instead of **deterministic inference**. Therefore, the goal of the learning task is to provide a **posterior distribution** $\mathcal{Q} \in \mathcal{M}(\mathcal{H})$ such that the following **expected risk** is minimized:

$$R_D(\mathcal{Q}) = \mathbb{E}_{h\sim\mathcal{Q}}(R_D(h)) = \mathbb{E}_{h\sim\mathcal{Q}}\mathbb{E}_{z\sim\mathcal{D}} l(h, z) \tag{17}$$

Similarly, by substituting the data distribution with the dataset, the **empirical risk** is defined by the following equation:

$$R_S(\mathcal{Q}) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{h\sim\mathcal{Q}} l(h, z_i) \tag{18}$$

It is important to note that since the true distribution of data is unknowable, the expected risk $R_D(\mathcal{Q})$ is generally difficult to calculate directly, while the empirical error serves as an **unbiased surrogate**. The PAC-Bayes framework provides inequalities relating expected risk and empirical risk. Define the KL divergence from $\mathcal{Q}$ to $\mathcal{P}$ as:

$$KL(\mathcal{Q}\|\mathcal{P}) = \mathbb{E}_{h\sim\mathcal{Q}} \log \frac{\mathcal{Q}(h)}{\mathcal{P}(h)} \tag{19}$$

Remark C.1. In fact, the expected risk tends to be any Loss you select; it does not have a fixed formulation. However, in PAC theory, KL is generally adopted to reduce the difficulty of theoretical analysis, though we do not strictly follow this convention.

Further considering that some errors in the generalization process cannot be effectively estimated, a statistical complexity term is generally used to characterize the corresponding generalization error. A specific example is given in Lemma C.2.

Lemma C.2. *Fix $\lambda > \frac{1}{2}$, and assume the loss function takes values within a range of length L. For any $\delta > 0, m \in \mathbb{N}, \mathcal{D}, \mathcal{P} \in \mathcal{M}$, then with probability at least $1-\delta$, the following inequality holds for $\mathcal{Q} \in \mathcal{M}$:*

$$R_D(\mathcal{Q}) \leq \frac{1}{1-\frac{1}{2\lambda}} \left(R_S(\mathcal{Q}) + \frac{\lambda L}{m} \left(KL(\mathcal{Q} \| \mathcal{P}) + \log \frac{1}{\delta} \right) \right) \tag{20}$$

The result of Lemma C.2 provides an upper bound on the true risk, summarized by the empirical risk and a complexity term. An important fact is that empirical risk is generally hard to improve, while decomposing the complexity term helps us understand issues arising during learning and further provides ideas for improvement. In auditing tasks, the non-equilibrium distribution of $\mathcal{P}$ is a crucial process guiding our improvement of MIA. So, how can we improve upon KL divergence? We first define some important statistical metrics.

In statistical learning theory and information theory, metrics measuring the difference between two probability distributions are usually formalized as statistical distances. Among them, f-Divergence is a widely used family of metrics defined by a convex function representing the difference between distributions.

Definition C.3 (f-Divergence)**.** Let $\mathcal{P}$ and $\mathcal{Q}$ be two probability distributions defined on a measurable space $\mathcal{X}$, with probability density functions $p(x)$ and $q(x)$, respectively. Given a convex function $f : \mathbb{R} \to \mathbb{R}$ satisfying $f(1) = 0$, the f-divergence from $\mathcal{P}$ to $\mathcal{Q}$ is defined as:

$$D_f(\mathcal{P} \| \mathcal{Q}) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)} \right) q(x)\, dx \tag{21}$$

Alternatively, utilizing the Radon-Nikodym derivative, it is expressed as:

$$D_f(\mathcal{P} \| \mathcal{Q}) = \int_{\mathcal{X}} f\left(\frac{d\mathcal{P}}{d\mathcal{Q}} \right) d\mathcal{Q} \tag{22}$$

Different selections of the function $f(x)$ correspond to different statistical distance metrics. The following lists several common f-divergences and their corresponding generating functions $f(x)$:

- **KL Divergence (Kullback-Leibler Divergence):**
$$f(x) = x \ln x \tag{23}$$
- **Reverse KL Divergence:**
$$f(x) = -\ln x \tag{24}$$
- **Hellinger Distance:**
$$f(x) = (\sqrt{x} - 1)^2 \quad \text{or} \quad f(x) = 2(1 - \sqrt{x}) \tag{25}$$
- **Total Variation Distance:**
$$f(x) = \frac{1}{2}|x - 1| \tag{26}$$
- **Pearson χ^2-Divergence:** Usually defined as $f(x) = (x-1)^2$, but may also appear in different literature in the following equivalent forms (differing only by constant or linear terms):
$$f(x) = (x-1)^2, \quad f(x) = x^2 - 1, \quad \text{or} \quad f(x) = x^2 - x \tag{27}$$
- **Reverse Pearson χ^2-Divergence:**
$$f(x) = \frac{1}{x} - 1 \quad \text{or} \quad f(x) = \frac{1}{x} - x \tag{28}$$
- **Jensen-Shannon Divergence (JS Divergence):**
$$f(x) = \frac{1}{2} \left[(x+1) \ln \left(\frac{2}{x+1} \right) + x \ln x \right] \tag{29}$$
- **L_1 Norm:**
$$f(x) = |x - 1| \tag{30}$$

Having introduced the above work, we can now proceed to the proof of Theorem 2.1.

C.1 Proof of Theorem 2.1

Theorem C.4 (PAC-Bayesian Generalization Bound with χ^2 Divergence)**.** *Let $\mathcal{H}$ be a hypothesis class and $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$ a loss function. Let D be a data distribution and $S = (z_1, \ldots, z_m) \sim D^m$ an i.i.d. sample. For any prior distribution $\mathcal{P}$ over $\mathcal{H}$ and any posterior distribution $\mathcal{Q}$ such that $\mathcal{Q} \ll \mathcal{P}$, with probability at least $1 - \delta$ over the draw of S, it holds that*

$$R_D(\mathcal{Q}) \le R_S(\mathcal{Q}) + \sqrt{\frac{2}{m\delta}\left(\chi^2(\mathcal{Q}\|\mathcal{P}) + 1\right)}, \tag{31}$$

where $R_D(\mathcal{Q}) = \mathbb{E}_{h\sim\mathcal{Q}}[R_D(h)]$, $R_S(\mathcal{Q}) = \mathbb{E}_{h\sim\mathcal{Q}}[R_S(h)]$, and the χ^2-divergence is defined as

$$\chi^2(\mathcal{Q}\|\mathcal{P}) = \int_{\mathcal{H}} \left(\frac{d\mathcal{Q}(h)}{d\mathcal{P}(h)} - 1\right)^2 d\mathcal{P}(h). \tag{32}$$

Proof. Let $S = (z_1, \ldots, z_m) \sim D^m$ be a training sample drawn independently from the distribution D. For any hypothesis $h \in \mathcal{H}$, define its true risk and empirical risk as

$$R_D(h) = \mathbb{E}_{z\sim D}[\ell(h, z)], \qquad R_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i). \tag{33}$$

We further define the generalization gap of a single hypothesis as

$$\Delta_S(h) = R_D(h) - R_S(h). \tag{34}$$

Since

$$\mathbb{E}_{S\sim D^m}[R_S(h)] = R_D(h), \tag{35}$$

we have, for any fixed $h \in \mathcal{H}$,

$$\mathbb{E}_{S\sim D^m}[\Delta_S(h)] = 0. \tag{36}$$

For a posterior distribution $\mathcal{Q}$, the corresponding Gibbs true risk and empirical risk are given by

$$R_D(\mathcal{Q}) = \mathbb{E}_{h\sim\mathcal{Q}}[R_D(h)], \qquad R_S(\mathcal{Q}) = \mathbb{E}_{h\sim\mathcal{Q}}[R_S(h)]. \tag{37}$$

Therefore,

$$R_D(\mathcal{Q}) - R_S(\mathcal{Q}) = \mathbb{E}_{h\sim\mathcal{Q}}[\Delta_S(h)]. \tag{38}$$

Assume that $\mathcal{Q} \ll \mathcal{P}$, and let

$$r(h) = \frac{d\mathcal{Q}}{d\mathcal{P}}(h) \tag{39}$$

denote the Radon–Nikodym derivative of $\mathcal{Q}$ with respect to $\mathcal{P}$. Then,

$$\mathbb{E}_{h\sim\mathcal{Q}}[\Delta_S(h)] = \int_{\mathcal{H}} \Delta_S(h) r(h)\, d\mathcal{P}(h) = \mathbb{E}_{h\sim\mathcal{P}}[r(h)\Delta_S(h)]. \tag{40}$$

By the Cauchy–Schwarz inequality, we obtain

$$\mathbb{E}_{h\sim\mathcal{P}}[r(h)\Delta_S(h)] \le |\mathbb{E}_{h\sim\mathcal{P}}[r(h)\Delta_S(h)]| \le \sqrt{\mathbb{E}_{h\sim\mathcal{P}}[r(h)^2]}\sqrt{\mathbb{E}_{h\sim\mathcal{P}}[\Delta_S(h)^2]}. \tag{41}$$

By the definition of the χ^2-divergence,

$$\chi^2(\mathcal{Q}\|\mathcal{P}) = \int_{\mathcal{H}} (r(h) - 1)^2\, d\mathcal{P}(h) = \mathbb{E}_{h\sim\mathcal{P}}[r(h)^2] - 1, \tag{42}$$

where we used

$$\mathbb{E}_{h\sim\mathcal{P}}[r(h)] = 1. \tag{43}$$

Hence,

$$\mathbb{E}_{h\sim\mathcal{P}}[r(h)^2] = \chi^2(\mathcal{Q}\|\mathcal{P}) + 1. \tag{44}$$

Let

$$Z_S = \mathbb{E}_{h\sim\mathcal{P}}[\Delta_S(h)^2]. \tag{45}$$

Then,

$$R_D(\mathcal{Q}) - R_S(\mathcal{Q}) \le \sqrt{\chi^2(\mathcal{Q}\|\mathcal{P}) + 1}\sqrt{Z_S}. \tag{46}$$

It remains to derive a high-probability upper bound for Z_S. Since the prior $\mathcal{P}$ is independent of the sample S, by Fubini's theorem we have

$$\mathbb{E}_{S\sim D^m}[Z_S] = \mathbb{E}_{h\sim\mathcal{P}}\mathbb{E}_{S\sim D^m}[\Delta_S(h)^2]. \tag{47}$$

For any fixed h, since $\mathbb{E}_S[\Delta_S(h)] = 0$, we have

$$\mathbb{E}_{S\sim D^m}[\Delta_S(h)^2] = \mathrm{Var}_{S\sim D^m}(R_S(h)). \tag{48}$$

Moreover,

$$R_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h, z_i). \tag{49}$$

Since $z_1, \ldots, z_m$ are independent and identically distributed,

$$\mathrm{Var}_{S\sim D^m}(R_S(h)) = \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}_{z_i\sim D}(\ell(h, z_i)). \tag{50}$$

By the bounded-variance assumption, there exists a constant $\sigma^2 > 0$ such that, for all $h \in \mathcal{H}$,

$$\mathrm{Var}_{z\sim D}(\ell(h, z)) \le \sigma^2. \tag{51}$$

Therefore,

$$\mathrm{Var}_{S\sim D^m}(R_S(h)) \le \frac{\sigma^2}{m}. \tag{52}$$

Consequently,

$$\mathbb{E}_{S\sim D^m}[Z_S] \le \frac{\sigma^2}{m}. \tag{53}$$

Since $Z_S \ge 0$, Markov's inequality gives

$$\Pr_{S\sim D^m}\left(Z_S \ge \frac{\sigma^2}{m\delta}\right) \le \frac{\mathbb{E}_S[Z_S]}{\sigma^2/(m\delta)} \le \delta. \tag{54}$$

Thus, with probability at least $1 - \delta$,

$$Z_S \le \frac{\sigma^2}{m\delta}. \tag{55}$$

Substituting this bound into the previous Cauchy–Schwarz inequality yields that, with probability at least $1 - \delta$, for all $\mathcal{Q} \ll \mathcal{P}$,

$$R_D(\mathcal{Q}) - R_S(\mathcal{Q}) \le \sqrt{\chi^2(\mathcal{Q}\|\mathcal{P}) + 1}\sqrt{\frac{\sigma^2}{m\delta}}. \tag{56}$$

Equivalently,

$$R_D(\mathcal{Q}) \le R_S(\mathcal{Q}) + \sqrt{\frac{\sigma^2}{m\delta}\left(\chi^2(\mathcal{Q}\|\mathcal{P}) + 1\right)}. \tag{57}$$

This completes the proof. □

C.2 Proof of Corollary 2.2

Corollary C.5 (Auditing Generalization Bound for Membership Inference)**.** *Let $\ell : \mathcal{H} \times \mathcal{Z} \to [0, 1]$ be the loss function of a membership inference attacker. Let $\mathcal{D}_f$ denote the true auditing distribution, and let $\mathcal{D}_t$ denote the training distribution used by the attacker. Assume $\mathcal{D}_t \ll \mathcal{D}_f$ and define the Rényi-∞ divergence as*

$$D_\infty(\mathcal{D}_t\|\mathcal{D}_f) = \log\left(\operatorname*{ess\,sup}_{z\sim\mathcal{D}_f}\frac{d\mathcal{D}_t}{d\mathcal{D}_f}(z)\right). \tag{58}$$

Let $S = (z_1, \ldots, z_m) \sim \mathcal{D}_t^m$ be an i.i.d. training sample. Then, for any prior $\mathcal{P}$ over $\mathcal{H}$ and any posterior $\mathcal{Q} \ll \mathcal{P}$, with probability at least $1 - \delta$ over S, the true auditing risk satisfies

$$R_{\mathcal{D}_f}(\mathcal{Q}) \le R_S(\mathcal{Q}) + \sqrt{\frac{2}{m\delta}\left(\chi^2(\mathcal{Q}\|\mathcal{P}) + 1\right)} + \sqrt{\frac{1}{2}D_\infty(\mathcal{D}_t\|\mathcal{D}_f)}. \tag{59}$$

Proof. Let $S = (z_1, \ldots, z_m) \sim \mathcal{F}_t^m$ be the training sample used to learn the MIA attacker. For any distribution $\mathcal{F}$, define

$$R_{\mathcal{F}}(\mathcal{Q}) = \mathbb{E}_{h\sim\mathcal{Q}}\mathbb{E}_{z\sim\mathcal{F}}[\ell(h, z)]. \tag{60}$$

We decompose the auditing risk as

$$R_{\mathcal{F}_u}(\mathcal{Q}) - R_S(\mathcal{Q}) = (R_{\mathcal{F}_t}(\mathcal{Q}) - R_S(\mathcal{Q})) + (R_{\mathcal{F}_u}(\mathcal{Q}) - R_{\mathcal{F}_t}(\mathcal{Q}))\,. \tag{61}$$

The first term is the standard PAC-Bayesian statistical error under the attacker training distribution $\mathcal{F}_t$. By applying the PAC-Bayesian bound with χ^2-divergence under a bounded-variance loss to $\mathcal{F}_t$, with probability at least $1 - \delta$ over the draw of $S \sim \mathcal{F}_t^m$, for all posteriors $\mathcal{Q} \ll \mathcal{P}$,

$$R_{\mathcal{F}_t}(\mathcal{Q}) \le R_S(\mathcal{Q}) + \sqrt{\frac{\sigma^2}{m\delta}\,(\chi^2(\mathcal{Q}\|\mathcal{P}) + 1)}. \tag{62}$$

It remains to control the second term, which measures the auditing error caused by the distribution shift from $\mathcal{F}_t$ to $\mathcal{F}_u$. Since the loss is bounded in $[0, 1]$, for any fixed hypothesis $h \in \mathcal{H}$,

$$|R_{\mathcal{F}_u}(h) - R_{\mathcal{F}_t}(h)| \le \mathrm{TV}(\mathcal{F}_t, \mathcal{F}_u). \tag{63}$$

Taking expectation over $h \sim \mathcal{Q}$ gives

$$R_{\mathcal{F}_u}(\mathcal{Q}) - R_{\mathcal{F}_t}(\mathcal{Q}) \le \mathrm{TV}(\mathcal{F}_t, \mathcal{F}_u). \tag{64}$$

We now bound the total variation distance by the Rényi divergence of order infinity. Let

$$\rho(z) = \frac{d\mathcal{F}_t}{d\mathcal{F}_u}(z) \tag{65}$$

be the density ratio between the attacker training distribution $\mathcal{F}_t$ and the audited distribution $\mathcal{F}_u$, and define

$$\varepsilon = D_\infty(\mathcal{F}_t\|\mathcal{F}_u) = \log\left(\operatorname*{ess\,sup}_z \rho(z)\right). \tag{66}$$

If $\varepsilon = \infty$, then the auditing error term on the right-hand side diverges, and the desired inequality holds trivially. This case indicates the absence of finite density-ratio control between $\mathcal{F}_t$ and $\mathcal{F}_u$: the attacker training distribution is not sufficiently covered by the audited distribution, or the two distributions have insufficient overlap for this divergence-based auditing guarantee to be finite. Therefore, it suffices to consider the nontrivial case where $\varepsilon < \infty$.

When $\varepsilon < \infty$, we have

$$\rho(z) \le e^{\varepsilon} \tag{67}$$

for $\mathcal{F}_u$-almost every z. Let

$$A = \{z : \rho(z) \ge 1\}\,. \tag{68}$$

By the variational characterization of total variation distance,

$$\mathrm{TV}(\mathcal{F}_t, \mathcal{F}_u) = \mathcal{F}_t(A) - \mathcal{F}_u(A). \tag{69}$$

Moreover, since $\rho(z) \le e^{\varepsilon}$, we have

$$\mathcal{F}_u(A) = \int_A d\mathcal{F}_u \ge e^{-\varepsilon}\int_A \rho(z)\, d\mathcal{F}_u(z) = e^{-\varepsilon}\mathcal{F}_t(A). \tag{70}$$

Therefore,

$$\mathrm{TV}(\mathcal{F}_t, \mathcal{F}_u) \le \left(1 - e^{-\varepsilon}\right)\mathcal{F}_t(A) \le 1 - e^{-\varepsilon}. \tag{71}$$

Using the elementary inequality

$$1 - e^{-\varepsilon} \le \sqrt{\frac{\varepsilon}{2}}, \qquad \varepsilon \ge 0, \tag{72}$$

we obtain

$$\mathrm{TV}(\mathcal{F}_t, \mathcal{F}_u) \le \sqrt{\frac{1}{2}D_\infty(\mathcal{F}_t\|\mathcal{F}_u)}. \tag{73}$$

Consequently,

$$R_{\mathcal{F}_u}(\mathcal{Q}) - R_{\mathcal{F}_t}(\mathcal{Q}) \leq \sqrt{\frac{1}{2} D_\infty(\mathcal{F}_t \| \mathcal{F}_u)}. \tag{74}$$

Combining the PAC-Bayesian statistical error under $\mathcal{F}_t$ with the auditing error bound yields that, with probability at least $1-\delta$, for all $\mathcal{Q} \ll \mathcal{P}$,

$$R_{\mathcal{F}_u}(\mathcal{Q}) \leq R_S(\mathcal{Q}) + \sqrt{\frac{\sigma^2}{m\delta}\left(\chi^2(\mathcal{Q}\|\mathcal{P}) + 1\right)} + \sqrt{\frac{1}{2} D_\infty(\mathcal{F}_t \| \mathcal{F}_u)}. \tag{75}$$

This completes the proof. □

C.3 Proof of Proposition 3.1

Proposition C.6. *If $\mathcal{F}_u$ is modeled as a mixture of $\mathcal{F}_n$ and $\mathcal{F}_m$ with proportions ρ and $1-\rho$, respectively, i.e.,*

$$\mathcal{F}_u = \rho \mathcal{F}_n + (1-\rho)\mathcal{F}_m, \tag{76}$$

then its mean μ_u and covariance Σ_u satisfy

$$\mu_u = \rho\mu_n + (1-\rho)\mu_m, \quad \Sigma_u = \rho\Sigma_n + (1-\rho)\Sigma_m + (\rho - \rho^2)(\mu_n - \mu_m)(\mu_n - \mu_m)^\top. \tag{77}$$

Proof of Proposition 3.1. Let $x \sim \mathcal{F}_u$ be a random variable drawn from the audited mixture distribution. Since

$$\mathcal{F}_u = \rho \mathcal{F}_n + (1-\rho)\mathcal{F}_m, \tag{78}$$

the mean of $\mathcal{F}_u$ is

$$\begin{aligned} \mu_u &= \mathbb{E}_{x\sim\mathcal{F}_u}[x] \\ &= \rho \mathbb{E}_{x\sim\mathcal{F}_n}[x] + (1-\rho)\mathbb{E}_{x\sim\mathcal{F}_m}[x] \\ &= \rho\mu_n + (1-\rho)\mu_m. \end{aligned} \tag{79}$$

Next, we derive the covariance of $\mathcal{F}_u$. By definition,

$$\begin{aligned} \Sigma_u &= \mathbb{E}_{x\sim\mathcal{F}_u}\left[(x-\mu_u)(x-\mu_u)^\top\right] \\ &= \rho\, \mathbb{E}_{x\sim\mathcal{F}_n}\left[(x-\mu_u)(x-\mu_u)^\top\right] + (1-\rho)\, \mathbb{E}_{x\sim\mathcal{F}_m}\left[(x-\mu_u)(x-\mu_u)^\top\right] \\ &= \rho\, \mathbb{E}_{x\sim\mathcal{F}_n}\left[(x-\mu_n+\mu_n-\mu_u)(x-\mu_n+\mu_n-\mu_u)^\top\right] \\ &\quad + (1-\rho)\, \mathbb{E}_{x\sim\mathcal{F}_m}\left[(x-\mu_m+\mu_m-\mu_u)(x-\mu_m+\mu_m-\mu_u)^\top\right] \\ &= \rho\Sigma_n + (1-\rho)\Sigma_m + \rho(\mu_n-\mu_u)(\mu_n-\mu_u)^\top + (1-\rho)(\mu_m-\mu_u)(\mu_m-\mu_u)^\top. \end{aligned} \tag{80}$$

Using

$$\mu_u = \rho\mu_n + (1-\rho)\mu_m, \tag{81}$$

we have

$$\mu_n - \mu_u = (1-\rho)(\mu_n - \mu_m), \qquad \mu_m - \mu_u = -\rho(\mu_n - \mu_m). \tag{82}$$

Substituting these identities into the covariance expression gives

$$\begin{aligned} \Sigma_u &= \rho\Sigma_n + (1-\rho)\Sigma_m + \rho(1-\rho)^2(\mu_n-\mu_m)(\mu_n-\mu_m)^\top \\ &\quad + (1-\rho)\rho^2(\mu_n-\mu_m)(\mu_n-\mu_m)^\top \\ &= \rho\Sigma_n + (1-\rho)\Sigma_m + \left[\rho(1-\rho)^2 + (1-\rho)\rho^2\right](\mu_n-\mu_m)(\mu_n-\mu_m)^\top \\ &= \rho\Sigma_n + (1-\rho)\Sigma_m + \rho(1-\rho)(\mu_n-\mu_m)(\mu_n-\mu_m)^\top \\ &= \rho\Sigma_n + (1-\rho)\Sigma_m + (\rho-\rho^2)(\mu_n-\mu_m)(\mu_n-\mu_m)^\top. \end{aligned} \tag{83}$$

This completes the proof. □

C.4 Proof of Lemma 3.2

Lemma C.7 (Solution of SMI). *Let $d = \mu_n - \mu_m, B = dd^\top, A = \Sigma_u - \Sigma_m, G = \Sigma_n - \Sigma_m + B$. SMI estimates the mixture ratio by solving:*

$$\min_{0\le\rho\le1} \left\|\Sigma_u - \left[\rho\Sigma_n + (1-\rho)\Sigma_m + (\rho-\rho^2)(\mu_n-\mu_m)(\mu_n-\mu_m)^\top\right]\right\|_F^2 \tag{84}$$

The solution is selected from the search space $\mathcal{C}$:

$$\rho^\star = \arg\min_{\rho\in\mathcal{C}} \left\|A - \rho G + \rho^2 B\right\|_F^2, \tag{85}$$

The search space $\mathcal{C}$ is:

$$\mathcal{C} = \{0,1\} \cup \left\{\rho\in[0,1] : 2\|B\|_F^2\rho^3 - 3\langle G,B\rangle_F\rho^2 + \left(\|G\|_F^2 + 2\langle A,B\rangle_F\right)\rho - \langle A,G\rangle_F = 0\right\}, \tag{86}$$

where, $\langle A,B\rangle_F = \mathrm{tr}(A^\top B)$ denotes the Frobenius inner product. Specifically, the solution process can be described as first solving a univariate cubic equation in ρ, and then substituting the obtained roots into the search space to select the minimizer.

Proof of Lemma 3.2. By Proposition 3.1, the covariance predicted by the mixture model with mixture ratio ρ is

$$\widehat{\Sigma}_u(\rho) = \rho\Sigma_n + (1-\rho)\Sigma_m + (\rho-\rho^2)(\mu_n-\mu_m)(\mu_n-\mu_m)^\top. \tag{87}$$

Let

$$d = \mu_n - \mu_m, \qquad B = dd^\top, \qquad A = \Sigma_u - \Sigma_m, \qquad G = \Sigma_n - \Sigma_m + B. \tag{88}$$

Then

$$\widehat{\Sigma}_u(\rho) = \Sigma_m + \rho(\Sigma_n - \Sigma_m + B) - \rho^2 B = \Sigma_m + \rho G - \rho^2 B. \tag{89}$$

Therefore, the covariance matching objective can be rewritten as

$$\left\|\Sigma_u - \widehat{\Sigma}_u(\rho)\right\|_F^2 = \left\|A - \rho G + \rho^2 B\right\|_F^2. \tag{90}$$

Thus, SMI estimates ρ by solving the one-dimensional constrained optimization problem

$$\rho^\star = \arg\min_{0\le\rho\le1} J(\rho), \qquad J(\rho) = \left\|A - \rho G + \rho^2 B\right\|_F^2. \tag{91}$$

We now derive the candidate set for the global minimizer. Using the Frobenius inner product

$$\langle X, Y\rangle_F = \mathrm{tr}(X^\top Y), \tag{92}$$

the objective can be expanded as

$$\begin{aligned} J(\rho) &= \left\langle A - \rho G + \rho^2 B, A - \rho G + \rho^2 B\right\rangle_F \\ &= \|A\|_F^2 - 2\rho\langle A,G\rangle_F + \rho^2\|G\|_F^2 + 2\rho^2\langle A,B\rangle_F \\ &\quad - 2\rho^3\langle G,B\rangle_F + \rho^4\|B\|_F^2. \end{aligned} \tag{93}$$

Since $J(\rho)$ is a continuous polynomial on the compact interval $[0,1]$, it attains a global minimizer. Any global minimizer must either lie at the boundary points $\rho = 0$ or $\rho = 1$, or be an interior stationary point satisfying $J'(\rho) = 0$.

Differentiating $J(\rho)$ gives

$$\begin{aligned} J'(\rho) &= -2\langle A,G\rangle_F + 2\rho\|G\|_F^2 + 4\rho\langle A,B\rangle_F \\ &\quad - 6\rho^2\langle G,B\rangle_F + 4\rho^3\|B\|_F^2. \end{aligned} \tag{94}$$

Dividing by 2, the stationary condition $J'(\rho) = 0$ is equivalent to

$$2\|B\|_F^2\rho^3 - 3\langle G,B\rangle_F\rho^2 + \left(\|G\|_F^2 + 2\langle A,B\rangle_F\right)\rho - \langle A,G\rangle_F = 0. \tag{95}$$

Therefore, all possible global minimizers are contained in the finite candidate set

$$\mathcal{C} = \{0,1\} \cup \left\{\rho\in[0,1] : 2\|B\|_F^2\rho^3 - 3\langle G,B\rangle_F\rho^2 + \left(\|G\|_F^2 + 2\langle A,B\rangle_F\right)\rho - \langle A,G\rangle_F = 0\right\}. \tag{96}$$

Consequently, the SMI estimate is obtained by evaluating the objective on this candidate set and selecting the minimizer:

$$\rho^\star = \arg\min_{\rho\in\mathcal{C}} \left\|A - \rho G + \rho^2 B\right\|_F^2. \tag{97}$$

This completes the proof. □

C.5 Proof of Lemma 3.3

Lemma C.8. *For the embedded $\mu_f^{(k)}, \mu_v^{(k)}, \mu_t^{(k)}$, the optimization problem (10) is equivalent to the following convex quadratic programming problem:*

$$\min_{0\le\alpha\le 1} \ \alpha^2\|\mu_v^{(k)} - \mu_t^{(k)}\|_{\mathcal{H}}^2 - 2\alpha\langle \mu_f^{(k)} - \mu_t^{(k)}, \mu_v^{(k)} - \mu_t^{(k)}\rangle_{\mathcal{H}} \tag{98}$$

This provides an efficient solution path for the optimization problem (**??**).

Proof. We begin by expanding the squared Hilbert space norm in the objective of problem (10). Recall that for any $a, b \in \mathcal{H}$, we have $\|a-b\|_{\mathcal{H}}^2 = \langle a-b, a-b\rangle_{\mathcal{H}}$. Applying this to the objective function:

$$\begin{aligned}&\|\mu_f^{(k)} - \alpha\mu_v^{(k)} - (1-\alpha)\mu_t^{(k)}\|_{\mathcal{H}}^2 \\ &= \big\langle \mu_f^{(k)} - \alpha\mu_v^{(k)} - (1-\alpha)\mu_t^{(k)},\ \mu_f^{(k)} - \alpha\mu_v^{(k)} - (1-\alpha)\mu_t^{(k)}\big\rangle_{\mathcal{H}}.\end{aligned}$$

We simplify the expression inside the inner product. Observe that:

$$\alpha\mu_v^{(k)} + (1-\alpha)\mu_t^{(k)} = \mu_t^{(k)} + \alpha(\mu_v^{(k)} - \mu_t^{(k)}). \tag{99}$$

Therefore,

$$\mu_f^{(k)} - \alpha\mu_v^{(k)} - (1-\alpha)\mu_t^{(k)} = (\mu_f^{(k)} - \mu_t^{(k)}) - \alpha(\mu_v^{(k)} - \mu_t^{(k)}). \tag{100}$$

Let us denote:

$$A := \mu_f^{(k)} - \mu_t^{(k)}, \quad B := \mu_v^{(k)} - \mu_t^{(k)}. \tag{101}$$

Then the objective becomes:

$$\|A - \alpha B\|_{\mathcal{H}}^2 = \langle A - \alpha B, A - \alpha B\rangle_{\mathcal{H}}. \tag{102}$$

Expanding the inner product using bilinearity and symmetry:

$$\begin{aligned}\|A - \alpha B\|_{\mathcal{H}}^2 &= \langle A, A\rangle_{\mathcal{H}} - 2\alpha\langle A, B\rangle_{\mathcal{H}} + \alpha^2\langle B, B\rangle_{\mathcal{H}} \\ &= \|A\|_{\mathcal{H}}^2 - 2\alpha\langle A, B\rangle_{\mathcal{H}} + \alpha^2\|B\|_{\mathcal{H}}^2.\end{aligned}$$

Since $\|A\|_{\mathcal{H}}^2$ is independent of α, it does not affect the minimizer of the optimization problem over $\alpha \in [0,1]$. Therefore, minimizing the original objective is equivalent to minimizing the function:

$$\alpha^2\|B\|_{\mathcal{H}}^2 - 2\alpha\langle A, B\rangle_{\mathcal{H}}. \tag{103}$$

Substituting back the definitions of A and B, we obtain:

$$\alpha^2\|\mu_v^{(k)} - \mu_t^{(k)}\|_{\mathcal{H}}^2 - 2\alpha\langle \mu_f^{(k)} - \mu_t^{(k)},\ \mu_v^{(k)} - \mu_t^{(k)}\rangle_{\mathcal{H}}, \tag{104}$$

which is precisely the objective in equation (13).

Finally, note that the objective in (13) is a quadratic function in α with a non-negative leading coefficient $\|\mu_v^{(k)} - \mu_t^{(k)}\|_{\mathcal{H}}^2 \ge 0$. Hence, it is convex in α, and the constraint $0 \le \alpha \le 1$ defines a compact convex set. Therefore, problem (13) is a convex quadratic programming problem, which admits a unique global minimizer that can be efficiently computed.

This completes the proof. □

D Theoretical Elaboration on SMI-M and MMD

In this section, we elucidate the motivation behind SMI-M. In traditional statistical inference and optimization problems, the similarity between two random variables or feature distributions is often measured by matching the first moment (mean) or the second moment (covariance matrix) of the samples. For instance, in the SMI-Σ method, the objective is to align the empirical covariance $\widehat{\Sigma}_f$ of the auditing features $f(\cdot; w)$ with the target covariance Σ^*. However, relying solely on low-order moments presents significant limitations:

- Low-order moments fail to characterize high-order structures of the distribution (such as skewness, kurtosis, and multimodality).
- In high-dimensional spaces, the estimation of covariance matrices is susceptible to noise and small sample sizes, leading to the "curse of dimensionality".
- If the true distribution is non-Gaussian, matching only the first two moments may completely overlook critical differences.

To overcome these issues, modern approaches tend to directly measure the **distance between entire probability distributions** rather than just their moments. Among these, the **Maximum Mean Discrepancy** (MMD) serves as a non-parametric distribution metric with a solid theoretical foundation and computational feasibility.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a non-empty set (typically a subset of Euclidean space), and let $\mathbb{P}$ and $\mathbb{Q}$ be two Borel probability measures defined on $\mathcal{X}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a **Mercer kernel function**, which satisfies the following conditions:

1. Symmetry: $k(x, x') = k(x', x)$ for all $x, x' \in \mathcal{X}$;
2. Positive Definiteness: For any finite set of points $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ and any real coefficients $\{a_i\}_{i=1}^n \subset \mathbb{R}$, the following holds:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0.$$

According to Mercer's theorem, there exists a unique **Reproducing Kernel Hilbert Space** (RKHS) $\mathcal{H}_k$, where the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ satisfies the **reproducing property**:

$$\langle f, k(x, \cdot) \rangle_{\mathcal{H}_k} = f(x), \quad \forall f \in \mathcal{H}_k,\, x \in \mathcal{X}.$$

Furthermore, the RKHS $\mathcal{H}_k$ can be constructed via a feature map $\phi : \mathcal{X} \to \mathcal{H}_k$, such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_k}$. This mapping embeds the input space into a (potentially infinite-dimensional) Hilbert space.

Given a probability measure $\mathbb{P}$, its **kernel mean embedding** in the RKHS $\mathcal{H}_k$ is defined as the Bochner integral:

$$\mu_{\mathbb{P}}^{(k)} := \mathbb{E}_{x \sim \mathbb{P}}[\phi(x)] = \int_{\mathcal{X}} \phi(x)\, d\mathbb{P}(x) \in \mathcal{H}_k,$$

provided that this expectation exists in $\mathcal{H}_k$ (e.g., this holds when $\mathbb{E}_{x \sim \mathbb{P}}[\sqrt{k(x, x)}] < \infty$). This embedding compresses the entire distribution $\mathbb{P}$ into a single vector within the RKHS, thereby transforming the problem of distribution comparison into a geometric problem within a Hilbert space.

The **Maximum Mean Discrepancy** (MMD) is defined as the Hilbert norm distance between the embeddings of the two distributions:

$$\mathrm{MMD}_k(\mathcal{P}, \mathcal{Q}) := \left\| \mu_{\mathcal{P}}^{(k)} - \mu_{\mathcal{Q}}^{(k)} \right\|_{\mathcal{H}_k}. \tag{105}$$

Utilizing the reproducing property of the RKHS and the linearity of the inner product, the square of MMD can be expanded into a form dependent solely on the expectation of the kernel function:

$$\begin{aligned} \mathrm{MMD}_k^2(\mathcal{P}, \mathcal{Q}) &= \left\langle \mu_{\mathcal{P}}^{(k)} - \mu_{\mathcal{Q}}^{(k)}, \mu_{\mathcal{P}}^{(k)} - \mu_{\mathcal{Q}}^{(k)} \right\rangle_{\mathcal{H}_k} \\ &= \mathcal{E}_{x, x' \sim \mathcal{P}}[k(x, x')] + \mathcal{E}_{y, y' \sim \mathcal{Q}}[k(y, y')] - 2\mathcal{E}_{x \sim \mathcal{P},\, y \sim \mathcal{Q}}[k(x, y)]. \end{aligned} \tag{106}$$

This expression demonstrates that MMD relies only on pairwise kernel values between samples, without the need to explicitly compute high-dimensional feature maps.

It is worth noting that whether MMD constitutes a valid **metric** depends on the properties of the kernel function k. If k is a **characteristic kernel**, satisfying

$$\mathrm{MMD}_k(\mathcal{P}, \mathcal{Q}) = 0 \iff \mathcal{P} = \mathcal{Q},$$

then MMD induces a metric on the space of probability measures. Sufficient conditions include: k is continuous, bounded, and the corresponding RKHS $\mathcal{H}_k$ is dense in $\mathcal{X}$ (e.g., the Gaussian RBF kernel satisfies this property on compact sets).

D.1 Empirical Estimation

In practical applications, assuming we only have independent and identically distributed (i.i.d.) samples $\{x_i\}_{i=1}^n$ from $\mathcal{P}$ and samples $\{y_j\}_{j=1}^m$ from $\mathcal{Q}$, MMD can be estimated via empirical mean embeddings. Define the empirical embeddings:

$$\hat{\mu}_{\mathcal{P}}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i), \quad \hat{\mu}_{\mathcal{Q}}^{(k)} = \frac{1}{m}\sum_{j=1}^{m}\phi(y_j).$$

The **biased estimator** is given by:

$$\widehat{\mathrm{MMD}}^2_{k,\text{biased}} = \left\|\hat{\mu}_{\mathcal{P}}^{(k)} - \hat{\mu}_{\mathcal{Q}}^{(k)}\right\|^2_{\mathcal{H}_k} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(x_i,x_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(y_i,y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}k(x_i,y_j). \tag{107}$$

This estimator possesses a systematic bias under finite samples but is consistent (i.e., it converges in probability to the true MMD as $n, m \to \infty$).

To eliminate the bias from auto-correlation terms, an **unbiased estimator** can be adopted:

$$\widehat{\mathrm{MMD}}^2_{k,\text{unbiased}} = \frac{1}{n(n-1)}\sum_{i\neq j}k(x_i,x_j) + \frac{1}{m(m-1)}\sum_{i\neq j}k(y_i,y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}k(x_i,y_j), \tag{108}$$

where the first and second terms exclude the diagonal entries where $i = j$. This estimator is unbiased when $n, m \geq 2$, although it may exhibit larger variance with small sample sizes.

D.2 Common Kernel Functions and Calculation Methods

The empirical performance of MMD depends on the choice of kernel function $k(x, x')$. For distance-based kernels, the *bandwidth* or scale parameter, usually denoted by $\sigma > 0$, controls how rapidly the kernel value decreases as the distance between two samples increases. A smaller bandwidth makes the kernel more sensitive to local differences, whereas a larger bandwidth produces smoother similarity scores. In practice, σ is commonly selected by the median heuristic or cross-validation.

Commonly used kernel functions can be computed as follows:

1. **Gaussian RBF Kernel**

$$k_{\mathrm{RBF}}(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right).$$

   To compute this kernel, first calculate the squared Euclidean distance

$$\|x-x'\|_2^2 = \sum_{i=1}^{d}(x_i - x_i')^2,$$

   then substitute it into the exponential function. The bandwidth σ determines the scale at which two samples are regarded as similar.

2. **Laplacian Kernel**

$$k_{\mathrm{Lap}}(x,x') = \exp\left(-\frac{\|x-x'\|_1}{\sigma}\right),$$

   where

$$\|x-x'\|_1 = \sum_{i=1}^{d}|x_i - x_i'|.$$

   The calculation first sums the absolute coordinate-wise differences, then divides the result by σ and applies the negative exponential.

3. **Linear Kernel**

$$k_{\text{lin}}(x, x') = x^\top x'.$$

This kernel is computed directly by taking the inner product between two vectors:

$$x^\top x' = \sum_{i=1}^{d} x_i x_i'.$$

4. **Polynomial Kernel**

$$k_{\text{poly}}(x, x') = (x^\top x' + c)^p,$$

where $c \geq 0$ is the bias term and $p \in \mathbb{N}$ is the polynomial degree. Its computation consists of first evaluating the inner product $x^\top x'$, adding the bias term c, and then raising the result to the power p.

5. **Cosine Kernel**

$$k_{\cos}(x, x') = \frac{x^\top x'}{\|x\|_2 \|x'\|_2}.$$

This kernel measures the cosine similarity between two nonzero vectors. It is computed by taking the inner product of x and x', then normalizing it by the product of their Euclidean norms.

After choosing a kernel, the empirical MMD estimator can be obtained by substituting sample pairs from the two distributions into the corresponding kernel function.

E The Complete Experimental

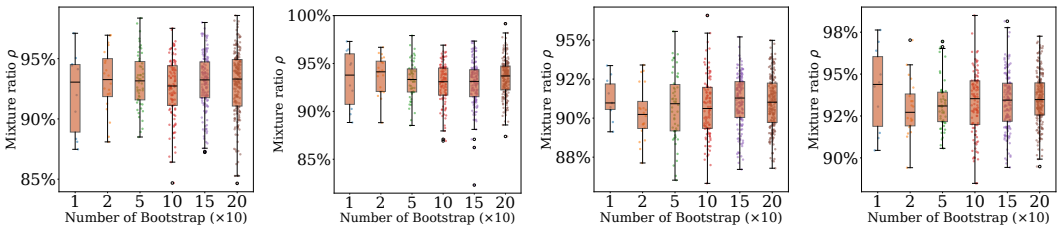
The complete experimental results are shown in Table 7.

Table 7: Performance auditing of SMI and MIA-based methods on the complete unlearning task. The Dataset column reports the model, task, unlearning percentage, and number of shadow models.Experimental results for full datasets, with the **best** and second best results are highlighted.

Dataset	Metric	SMI	SMI-M	RULI	IAM	RMIA	EMIA	LiRA	Unleak	Ramdom
Resnet18-Cifar10-5%-64	FNR	–	–	13.28%	12.72%	49.68%	70.82%	13.80%	41.24%	49.52%
	FPR	–	–	0.00%	13.08%	50.04%	70.72%	0.00%	43.48%	50.88%
	ρ^*	88.89%	**92.44%**	87.94%	88.67%	66.68%	39.10%	87.84%	62.58%	50.06%
Resnet18-Cifar10-10%-64	FNR	–	–	19.38%	14.22%	56.98%	61.22%	18.20%	52.44%	49.88%
	FPR	–	–	0.00%	18.28%	42.24%	67.83%	0.00%	31.88%	50.17%
	ρ^*	87.70%	**91.80%**	83.24%	83.37%	61.68%	30.90%	89.94%	53.28%	50.02%
Resnet18-Cifar100-5%-16	FNR	–	–	15.60%	17.76%	87.80%	24.72%	10.47%	81.04%	48.72%
	FPR	–	–	0.00%	17.96%	87.96%	22.24%	0.00%	82.88%	49.52%
	ρ^*	79.64%	**89.84%**	83.00%	89.36%	67.19%	84.64%	86.64%	28.52%	50.88%
Resnet18-Cifar100-10%-16	FNR	–	–	31.00%	54.88%	58.96%	87.04%	32.24%	62.84%	49.71%
	FPR	–	–	0.00%	52.12%	60.92%	87.00%	0.00%	64.36%	51.02%
	ρ^*	77.56%	**91.84%**	87.48%	84.76%	81.74%	51.46%	88.94%	36.90%	50.58%
Resnet50-Cifar100-5%-32	FNR	–	–	36.76%	86.72%	91.92%	66.52%	37.41%	93.08%	47.68%
	FPR	–	–	0.00%	63.83%	56.00%	57.56%	0.00%	47.69%	50.15%
	ρ^*	81.22%	**92.64%**	73.08%	63.56%	58.24%	67.84%	87.04%	10.40%	54.68%
Resnet50-Cifar100-10%-32	FNR	–	–	26.90%	87.73%	91.50%	70.72%	38.29%	93.56%	49.71%
	FPR	–	–	0.00%	61.69%	53.21%	57.81%	0.00%	49.77%	42.27%
	ρ^*	83.46%	**93.95%**	74.44%	64.98%	60.74%	69.94%	87.38%	4.68%	54.14%
Resnet18-Cinic10-5%-16	FNR	–	–	28.44%	82.32%	90.64%	9.24%	5.36%	14.96%	49.92%
	FPR	–	–	0.00%	81.68%	90.20%	8.36%	0.00%	14.52%	51.13%
	ρ^*	89.12%	**92.10%**	88.47%	37.53%	29.44%	88.24%	89.27%	87.28%	50.05%
Resnet18-Cinic10-10%-16	FNR	–	–	24.17%	81.44%	94.08%	85.56%	17.64%	29.92%	49.52%
	FPR	–	–	0.00%	76.84%	92.56%	78.84%	0%	22.04%	50.88%
	ρ^*	87.24%	**92.41%**	91.15%	40.34%	24.98%	35.88%	89.62%	78.71%	49.83%
Resnet50-Cinic10-5%-16	FNR	–	–	11.68%	82.32%	90.64%	15.24%	11.36%	14.96%	49.24%
	FPR	–	–	0.00%	81.68%	90.20%	15.36%	0.00%	14.52%	50.21%
	ρ^*	89.31%	**90.72%**	89.74%	37.53%	29.44%	88.26%	89.92%	87.28%	49.36%
Resnet50-Cinic10-10%-16	FNR	–	–	11.36%	29.92%	11.88%	22.08%	11.18%	19.46%	49.82%
	FPR	–	–	0.00%	30.96%	11.28%	21.68%	0.00%	14.04%	50.23%
	ρ^*	90.89%	**92.27%**	89.88%	85.35%	87.64%	89.66%	90.47%	78.02%	49.76%
VIT-Cinic10-5%-16	FNR	–	–	15.68%	46.96%	15.56%	36.60%	19.27%	81.28%	49.42%
	FPR	–	–	0.00%	48.16%	14.72%	35.36%	0.00%	81.96%	50.48%
	ρ^*	89.72%	**92.13%**	85.05%	53.62%	85.30%	65.52%	78.62%	18.84%	50.52%
VIT-Cinic10-10%-16	FNR	–	–	15.92%	57.40%	85.04%	65.36%	18.76%	43.24%	49.97%
	FPR	–	–	0.00%	53.37%	86.12%	65.24%	0.00%	44.76%	51.93%
	ρ^*	91.44%	**93.62%**	89.84%	41.01%	15.33%	36.66%	83.51%	55.71%	51.04%

E.1 Robustness Experiments

We provide additional visualizations of the experiments in Figures 5 and 6.



(a) ResNet18-5%-SMI (b) ResNet18-5%-SMI-M (c) ResNet18-10%-SMI (d) ResNet18-10%-SMI-M

Figure 5: Performance fluctuation of SMI and SMI-M with different numbers of resampling iterations on 5% and 10% randomly unlearned data.

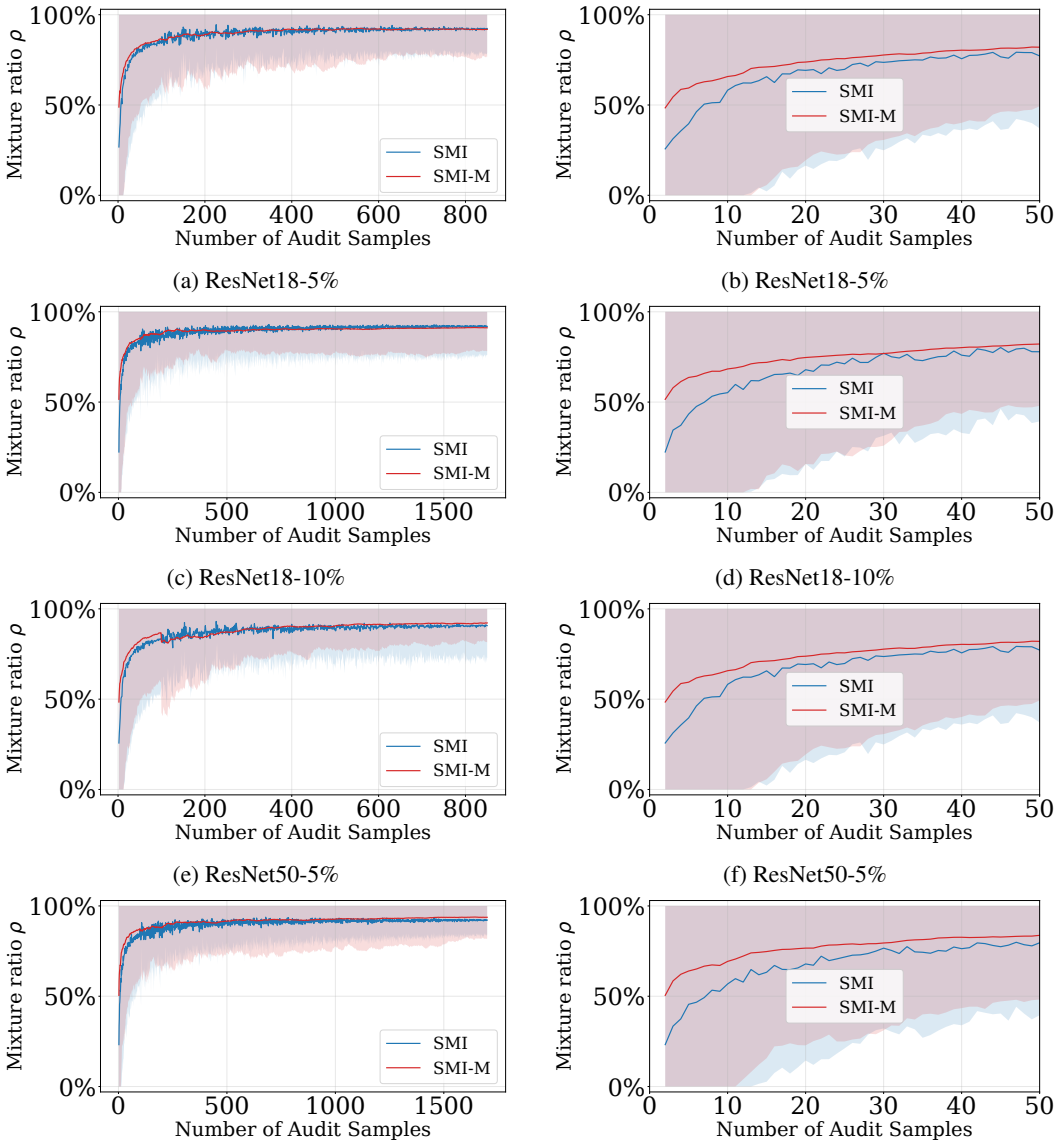


Figure 6: Performance fluctuation of SMI and SMI-M with different numbers of resampling iterations on 5% and 10% randomly unlearned data.

E.2 Experimental Environment

We conduct the experiments on eight NVIDIA RTX A6000 GPUs, with Python 3.11 as the experimental environment.

E.3 Unlearning methods

We select Retrain, FineTune, Fisher, and Scrub as baseline methods for machine unlearning, in order to compare different unlearning strategies in terms of forgetting effectiveness and retained performance. Let D_u denote the dataset to be unlearned, and let D_m denote the member dataset after removing D_u.

Retrain is the most direct unlearning baseline. It removes the unlearning dataset D_u from the original training set and trains a model with the same architecture from scratch using only the member dataset D_m. Since this model never observes D_u during training, Retrain is usually regarded as the ideal reference for machine unlearning. However, it requires a full retraining process and therefore incurs

substantial computational cost, so it is mainly used to evaluate the effectiveness of other unlearning methods.

FineTune adopts a fine-tuning-based unlearning strategy. Starting from the original model, it continues training the model using only the member dataset D_m, so that the model gradually shifts toward a state that relies more on the retained data. This method is simple to implement, has lower computational cost than retraining, and can weaken the influence of the unlearning data D_u to some extent. However, since the model parameters are inherited from the original model, FineTune cannot guarantee complete removal of the information contained in D_u.

Fisher adopts a Fisher-information-based unlearning method. It estimates the importance of model parameters to determine which parameters are more critical for model predictions, and then adjusts or perturbs the parameters accordingly to weaken the influence of the unlearning data D_u. The core idea of this method is to exploit local second-order information in the parameter space, so as to forget D_u while preserving the model performance on D_m as much as possible. Compared with simple fine-tuning, Fisher pays more attention to the differences in how different parameters contribute to the model behavior.

Scrub adopts a distillation-based unlearning method. It guides the model to deviate from the original predictions on the unlearning dataset D_u, while preserving the prediction behavior of the original model on the member dataset D_m. In this way, Scrub balances forgetting effectiveness and retained performance: it weakens the model's memorization of the target unlearning samples while minimizing performance degradation on the remaining data.

E.4 Relationship with Unlearning Inversion

SMI characterizes the membership-related behavior of forgotten samples in the auditing feature space, providing a perspective on unlearning inversion attacks. Intuitively, an auditing score close to 1 indicates more non-member-like behavior, which may make reconstruction harder. A lower score suggests stronger member-like characteristics and potentially greater vulnerability to inversion attacks.

We preliminarily examine this relationship by comparing the auditing scores and inversion mean squared errors of Fisher and Scrub, as shown in Table 8.

Table 8: The auditing scores and inversion mean squared errors are compared between Fisher and Scrub. Larger mean squared errors indicate greater reconstruction errors.

Unlearning method	Auditing score	Inversion MSE
Fisher	32.51%	0.19
Scrub	76.34%	0.33

Scrub yields both a higher auditing score and a larger reconstruction error than Fisher. This observation is consistent with the intuition that higher auditing scores may accompany more difficult reconstruction. However, these results provide only preliminary empirical support; the statistical relationship between auditing scores and inversion errors remains insufficiently characterized.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Justification: The abstract and introduction accurately reflect the paper's theoretical contributions, experimental scope, and main findings. The corresponding assumptions, analysis, and empirical evidence are provided in Sections and Appendices.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The main limitation of our SMI-based auditing method is its difficulty in auditing small-sample tasks. Because the method reduces reliance on external information sources, its performance degrades in settings with fewer than 20 samples per class.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: From the perspective of theoretical analysis, we follow the standard assumptions of PAC-Bayes. Apart from PAC-Bayes, our stochastic analysis does not require introducing additional assumptions, and all proofs are clearly organized in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our proposed algorithm is relatively simple and easy to reproduce. In contrast, the MIA baseline is more difficult to reproduce because shadow model training involves substantial randomness. For example, in the case of RuLI, the trained shadow models occupy more than 68 GB of storage, making it difficult for us to release such large files under anonymity constraints. Therefore, we will open-source them after the paper is accepted.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will open-source the code after the paper is accepted. Moreover, the experiments in the main paper are very simple and easy to reproduce; reviewers should be able to reproduce them from scratch within 24 hours.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide a detailed description of our experimental settings in the experimental setup section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We clearly present the results of repeated experiments using box plots and shaded line plots.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided the relevant details in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We use only standard public datasets and public model tasks, and our work strictly complies with the relevant ethical requirements.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: This paper aims to advance the successful auditing of machine unlearning. The SMI paradigm we propose facilitates better identification of whether samples are truly forgotten during the machine unlearning process, as well as detecting if false unlearning information is reported. This work is intended to promote the development of auditing tools in the field of privacy to achieve trustworthy artificial intelligence models.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This issue is not applicable to our work.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: This issue is not applicable to our work.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This issue is not applicable to our work.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This issue is not applicable to our work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This issue is not applicable to our work.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This issue is not applicable to our work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.