RMR: A Relative Membership Risk Measure for Machine Learning Models

Li Bai, Haibo Hu, Qingqing Ye, Jianliang Xu, Jin Li, Chengfang Fang, Jie Shi

***Abstract*—Privacy leakage poses a significant threat when machine learning foundation models trained on private data are released. One such threat is membership inference attacks (MIA), which determine whether a specific example was included in a model's training set. This paper shifts focus from developing new MIA algorithms to measuring a model's risk under MIA. We introduce a novel metric, Relative Membership Risk (RMR), which assesses a model's MIA vulnerability from a comparative standpoint. RMR calculates the difference in prediction loss for training examples relative to a predefined reference model, enabling risk comparison across models without needing to delve into details like training strategy, architecture, or data distribution. We also explore the selection of the reference model and show that using a high-risk reference model enhances the accuracy of the RMR measure. To identify the most vulnerable reference model, we propose an efficient iterative algorithm that selects the optimal model from a set of candidates. Through extensive empirical evaluations on various datasets and network architectures, we demonstrate that RMR is an accurate and efficient tool for measuring the membership privacy risk of both individual training examples and the overall machine learning model.**

***Index Terms*—Machine learning, membership inference attack, privacy leakage**

I. Introduction

THANKS to the rapid advancements in machine learning and knowledge discovery, companies like Google [1], Amazon [2], and Microsoft [3] now offer Machine Learning as a Service (MLaaS). These platforms enable users to train and deploy ML models using private datasets through web interfaces or out-of-the-box APIs. However, such models are vulnerable to exposing private information from their training data. For instance, Carlini *et al.* [4] demonstrated how a downstream task of GPT-2, namely autocompletion, could be exploited to successfully extract sensitive information such as an individual's full name, physical address, email address, phone number, and fax number. As a result, users often aim to release a safer model when a set of candidates is provided, relying on privacy risk assessments to mitigate potential privacy breaches.

Among the attacks that pose significant risks to data owners' privacy, membership inference attacks (MIAs) — which determine whether a query example is part of a model's training dataset — is a fundamental cause of privacy breaches [5], [6]. The first study was proposed by Shokri *et al.* [5] using shadow training. Since then, a range of related works have extended the scope and capability of MIA in generative models [7]–[9], regression models [10] and pre-trained models [11], [12]. These attacks pose a significant threat to an individual's data used for training ML models. For instance, if an attacker discovers that Alice's medical record was used to train an AIDS prediction model, they could infer that Alice is likely diagnosed with AIDS, resulting in a severe violation of her privacy. Beyond direct privacy violations, MIAs also serve as a foundation for more advanced inference attacks, such as model stealing [13] and model inversion [14], enhancing their effectiveness and extending their impact. As the reliance on sensitive data to train machine learning models continues to grow, assessing membership privacy risks has become a crucial aspect of privacy protection.

The work focuses on evaluating membership risk across multiple model candidates and selecting a privacy-preserving model for model owners. Such scenarios commonly arise in real-world scenarios, as organizations often develop or acquire multiple machine learning models to solve the same task. These models may vary in their architecture, training setting, or development approach. For instance, an e-commerce platform might develop various recommendation algorithms, such as neural networks and collaborative filtering, to assess which model best balances user experience with privacy protection, minimizing the risk of exposing sensitive user preferences [15]–[17]. In federated learning, multiple local models are trained by different participants (e.g., hospitals or banks) and later aggregated into a global model [18], [19]. Evaluating privacy risks is crucial in this context to ensure the selected global model is resilient to inference attacks.

A straightforward approach to evaluating membership risk is to measure the success rate of existing MIAs [20], [21]. However, this approach is time-consuming when applied to numerous candidate models and may underestimate the privacy risks posed by stronger attacks. For example, [21] evaluates the privacy risk score of training examples using a modified confidence-based MIA. Similarly, ML-DOCTOR [20] estimates a model's membership risk using a shadow training-based attack [5]. The computational cost becomes even more prohibitive when multiple attacks are conducted, and relying on a single attack method may fail to capture the full membership risk exposed by more powerful attacks [22]. Thus, an ideal membership privacy risk metric should focus on the root cause of MIAs, remain independent of any particular attack method, and be adaptable to future MIA techniques [23].

Li Bai, Haibo Hu (corresponding author), and Qingqing Ye are with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. Email: baili.bai@connect.polyu.hk, haibo.hu@polyu.edu.hk, qqing.ye@polyu.edu.hk. Jianliang Xu is with the Department of Computer Science, Hong Kong Baptist University. Email: xujl@comp.hkbu.edu.hk. Jin Li is with the School of Computer Science, Guangzhou University. Email: lijin@gzhu.edu.cn. Chengfang Fang and Jie Shi are with Huawei International. Email: fang.chengfang@huawei.com, shi.jie1@huawei.com.

Manuscript received Oct 31, 2023;

Previous research proposes attack-agnostic solutions to address this problem, though with certain limitations. These approaches are based on the underlying causes of membership privacy leakage, such as overfitting [5], [24] and model stability [24], [25]. For instance, overfitting is not a prerequisite for MIA. It estimates the model risk according to the generalization gap between training and test data, ignoring the different effects of training examples. Even in a well-generalized model, an adversary can identify vulnerable training examples with high precision [26]. Model stability, which measures the influence of a particular training example on the model output, often leads to a high computational cost for the leave-one-out process [23], [25].

In this paper, we propose an efficient and accurate MIA risk measure for machine learning models. The task of measuring the MIA risk of a target model is overcomplicated in practice as it is affected by many factors, e.g., model type [27], distribution of training data [5], [26], training strategy [24], [28], [29] and model memorization [30], [31]. Instead of measuring the membership privacy risk directly, we propose to measure Relative Membership Risk (RMR) of a model with respect to a *reference model*, which is acted by a trained model in the candidates without a retraining process. We treat the risk of a reference model as a *benchmark* and then derive the relative risk of target models on top of it. RMR is provable and practical to reveal the membership privacy of a training example or a machine learning model, which does not depend on specific training strategies, for example, combined with differential privacy mechanisms [32]. Furthermore, we show that a high-risk reference model should be chosen to measure RMR accurately among a set of candidate models to assess. Therefore, we propose an iterative algorithm to find the most vulnerable model among those models efficiently. Besides, RMR can evaluate the privacy risk of an individual example to identify those vulnerable to membership attacks.

In summary, our contributions are as follows:

1) We formally define a relative membership risk measure, namely RMR, to assess the membership inference risk of a training example or a model without depending on particular training strategies or data distribution beyond the example loss.
2) We analyze the impact of the risk of a reference model itself on the accuracy of RMR and devise an iterative algorithm to choose a suitable reference model in practice.
3) We conduct extensive experiments to demonstrate that RMR can reflect the practical MIA performance on various datasets and models. We also show that the MIA threats posed to a model can be mitigated by a risky-example-removal strategy based on the RMR measure.

The rest of this paper is organized as follows. Section II reviews MIAs and risk assessment. Section III proposes the RMR measure, and Section IV studies the reference model selection problem. The experimental results are shown in Section V. Finally, this paper is concluded in Section VI.

II. Related Work

This section reviews membership inference attacks in machine learning, followed by risk assessment against such attacks.

A. Membership Inference Attacks

A membership inference attack is a fundamental privacy threat in machine learning, aiming to deduce if a specific example is part of the training set by analyzing the model's output. This breach of privacy not only compromises the training dataset but also acts as a precursor to other inference attacks, including model extraction attack [13]. Additionally, the vulnerability to membership privacy breaches can be exacerbated by poisoning attacks [33], [34].

As for classification tasks, MIAs can be divided into two types: training-based and metric-based. Shokri *et al.* [5] first developed a training-based attack model to distinguish the membership status of a query sample according to its output from the target model. They trained shadow models to mimic the target model and then built a membership dataset for the attack model. Salem *et al.* [35] relaxed the assumptions of multiple shadow models and auxiliary data of the same distribution, and still achieved similar performance. Metric-based MIAs often set a threshold on a specific metric, such as prediction loss [24], [36], to infer membership status without training an attack model [21], [37]. Research works [38], [39] showed these attacks are even feasible for label-only scenarios where the full confidence scores are unavailable to attackers.

Besides attacks against classifiers, recent work has extended MIA to other ML models, such as generative and pre-trained models. For example, Heyas *et al.* [8] designed the first MIA on generative models in white-box and black-box settings. It detects the membership status by relying on the capacity of the discriminator. A concurrent study [9] investigated attacks against GAN and VAE by Monte Carlo approximation and reconstruction, respectively. For popular pre-trained models, Song *et al.* [40] leveraged the similarity score of embeddings to compromise membership status. The work [41] revealed that even without exposing the embedding layer, membership leakage persists on major language models.

As for defense, several countermeasures have been proposed to alleviate the threat of MIAs. For example, regularization techniques [5], [11], [38] and confidence score perturbation [5], [39], [42] are widely used to prevent membership inference attacks. Besides, complex mechanisms, such as differential privacy [8], [11], [20], [38], [42]–[45], adversarial learning [46], and knowledge distillation [47], have also been utilized to guard against privacy leakage of MIAs.

B. Membership Privacy Risk Measurement

Given the privacy leakage risks posed by MIAs, various approaches have been developed to measure its impact on machine learning models. These approaches are broadly classified into attack-related methods, which evaluate risk based on specific attacks, and attack-agnostic methods, which focus on the underlying causes of membership privacy leakage.

Attack-related methods, which evaluate risk based on the success rate of existing MIAs, provide a straightforward approach to measuring membership privacy risk [20], [21]. However, these methods are time-intensive when applied to multiple candidate models and may underestimate the leakage risks. For instance, [21] employs a modified confidence-based MIA to calculate the privacy risk score of training examples. Similarly, ML-DOCTOR [20] utilizes a shadow training-based attack [5] to estimate a model's membership risk. Such risk measurement approaches are heavily influenced by adversary knowledge, making them less reliable under varying conditions. Additionally, the computational cost escalates significantly when multiple attacks are performed. Moreover, relying on a single attack method may fail to capture the full extent of membership risk exposed by advanced attacks.

To keep independent of specific MIAs and adaptable to future attacks, attack-agnostic methods are designed to address membership leakage at its core. Existing works on this topic have explored the problem from two perspectives: the generalization gap and model stability. The generalization gap, also known as *overfitting*, is a sufficient but not necessary condition for a successful MIA [24]. Due to overparameterized networks and redundant training epochs, machine learning models often overfit and memorize training examples. While the connection between overfitting and MIA has been demonstrated both empirically [5], [35], [48], [49] and theoretically [24], recent studies have shown that certain MIAs can still succeed even when overfitting is not severe [4], [26]. Another limitation of overfitting as a measure of MIA risk is that it cannot measure the risk of a single training example.

Attack-agnostic approaches based on model stability focus on how much a training example influences the model's output. Examples that significantly affect the model are believed to pose a higher risk of membership inference breaches. Long *et al.* [25] quantitatively assesses the sensitivity of each training example and used the most influential examples as an empirical MIA risk measure, similar to the leave-one-out technique. However, this approach requires training a shadow model for each example and is only feasible for small training sets. To address the computational burden of the naive leave-one-out strategy, SHAPr [23] is proposed as an approximation method based on Shapley values. Model stability measures have received particular attention in models trained with differential privacy mechanisms [32], [50], which limit the change in a model's response to a specific example. [24] formally derives a bound for membership advantage in a differential private model, given by $e^{\epsilon} - 1$. Bernau *et al.* [28] obtain a tighter identifiability bound using Bayesian posterior belief. However, a range of works [24], [28], [51] show that these theoretical bounds are too loose to effectively measure membership risk in practice. Additionally, not all machine learning models, particularly those used in sensitive areas like healthcare or finance, can tolerate significant utility loss or be trained with differential privacy mechanisms.

Compared to existing approaches, our study aims to provide an effective, accurate, and attack-agnostic solution for estimating membership privacy risks for both training data records and models.

III. Privacy Risk under Membership Inference Attacks

This section presents the measure of privacy risk under membership inference attacks. The reason why MIAs work involves too many factors, including model architecture [27], training strategy [26], [52], data distribution [53] and overfitting of target model [24]. Hence it is intractable to consider all these factors and measure the membership risk directly. In this section, we take an alternative approach and define the risk in a relative manner. We first derive the absolute membership risk as the probability of an example being a member. Then we formally present the relative risk measure RMR to assess the MIA risk of a model.

TABLE I
Notations.

Notation	Description	Definition
D	Training dataset	$D = \{(x_i, y_i)\}_{i=1}^{N}$
f	Neural network	$f(x;\theta)$
(x, y)	Data example	-
m	Membership status	$m = \mathbb{I}(x \in D)$
θ	Model parameters	$\theta = \{\theta_i\}$
ℓ	Loss function	$\ell(y, p) = -\sum_{i=1}^{C} y_i \log p_i$
θ_t	Target model	$f(x;\theta_t)$
θ_r	Reference model	$f(x;\theta_r)$
σ	Sigmoid function	$\sigma(z) = \frac{1}{1+e^{-z}}$

A. Absolute Membership Risk

A supervised ML model learns from a labeled training set and makes predictions on unlabeled inputs [54]. Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote the training set, where N is the total number of examples, and (x_i, y_i) is the i-th sample with a feature vector x_i labeled by y_i. We treat an ML model as a mapping function $f(x;\theta)$, where x is an input example and θ is model parameters. Given a training set D, the model is trained to minimize a loss function $\ell(y, f(x;\theta))$, or simply $\ell(x, \theta)$ over D. Table I summarizes the notations used in this paper.

Membership inference attacks aim to deduce from a model an example's existence in the training set. Consequently, the probability of an example being a member denotes the risk of private information leakage, referring to how much the amount of exposure to attackers. As such, we formulate the absolute membership risk (AMR) as follows.

Definition 1 (absolute membership risk)**.** *Given a target model θ_t trained on D and a training example x, the absolute privacy leakage risk of θ_t about x is:*

$$amr(x, \theta_t) = P(m = 1 | x, \theta_t, D), \tag{1}$$

where m denotes the membership status that is 1 for a member, otherwise 0.

The absolute membership risk is essentially the posterior probability distribution over a target model and an example. Inspired by [36], the absolute membership risk can be further expressed as:

$$P(m = 1 | x, \theta_t, D) = \sigma\left[-\ell(x, \theta_t) + \tau_p(x) + t_\lambda\right], \tag{2}$$

where

$$\tau_p(x) = -\log\left(\int_{\theta'} \exp\left(-\ell\left(x,\theta'\right)\right) P\left(\theta'|D\backslash\{x\}\right) \mathrm{d}\theta'\right), \quad (3)$$

$$t_\lambda = \log\left(\frac{P(m=1)}{P(m=0)}\right).$$

The term $P\left(\theta'|D\backslash\{x\}\right)$ in Eq.(3) denotes the posterior distribution in which the example x is removed from D, the term t_λ denotes the log ratio of prior distribution of members ($P(m=1)$) to that of non-members ($P(m=0)$), and the term σ is the sigmoid function, that is, $\sigma(x) = 1/(1+e^{-x})$. As we take a further look at Eq.(2), the first two terms dominate the estimated membership probability, as the last term is a constant and approaches 0 if the query example has similar probabilities of being a member and a non-member. Using Jensen's inequality and the optimum over parameter space [55], Eq.(2) can be approximated as follows:

$$P(m=1|x,\theta_t,D) \approx \sigma\left[-\ell\left(x,\theta_t\right) + \ell\left(x,\theta'\right)\right], \quad (4)$$

where θ' denotes an *auxiliary model* that is trained on the training set $D\backslash\{x\}$. To mitigate the impact of factors unrelated to the specific example x, θ' should have the same training settings as θ_t.

The challenge remains in deriving the auxiliary model for the whole training set, although Eq.(4) provides a practical solution to compute the absolute membership risk of a single instance. When measuring the membership risk of a model, we need to consider the privacy leakage risks of all examples. One possible approach to achieve this is to adopt the leave-one-out strategy for each example in the training set, similar to [25], which evidently incurs significant computational costs. Another solution [37] is to introduce a shadow dataset drawn from the same data distribution as the training set.

Our methodology proposed in the following subsection measures the relative risk against a reference model, thus canceling out the auxiliary model and circumventing the heavy computation. It requires an individual reference model for the whole training dataset and avoids introducing an additional dataset when measuring the model risk. When dealing with a set of candidate models, it avoids the need for retraining by selecting one model to serve as the reference.

B. Relative Membership Risk

In this subsection, we present the definition of Relative Membership Risk (RMR), a relative MIA risk measure to the aforementioned membership inference. A model's risk is relative to a preset *reference model* trained on the same dataset. In essence, RMR designates the additional privacy leakage of a target model with respect to the reference model. Formally,

Definition 2 (relative membership risk)**.** *Given a reference model θ_r, a target model θ_t and a training example x, the relative privacy leakage risk of θ_t about x compared with θ_r under membership inference attacks is the difference of their absolute membership risks:*

$$rmr_{\theta_r}(x,\theta_t) = amr\left(x,\theta_t\right) - amr\left(x,\theta_r\right). \quad (5)$$

The following theorem provides an upper bound of RMR based on our discussion in Section III-A.

Theorem 1. *Given a target model θ_t and a reference model θ_r, if $amr\left(x,\theta_t\right) \leq amr\left(x,\theta_r\right)$, the rmr of an example x with θ_t with respect to θ_r is bounded as:*

$$rmr_{\theta_r}(x,\theta_t) \leq k\sigma\left[-\ell\left(x,\theta_t\right) + \ell\left(x,\theta_r\right)\right] - 1, \quad (6)$$

where $k = 2+e^C+e^{2C}$, and C is the upper bound of training example loss.

Proof. According to Eq.(4), we introduce an auxiliary model θ' that is trained on the same training set but without the sample x, then the relative risk is:

$$rmr_{\theta_r}(x,\theta_t) = \sigma\left[-\ell\left(x,\theta_t\right) + \ell\left(x,\theta'\right)\right] - \sigma\left[-\ell\left(x,\theta_r\right) + \ell\left(x,\theta'\right)\right].$$

According to Lemma 1, we have $\ell\left(x,\theta_t\right) \leq \ell\left(x,\theta'\right)$. On the assumption that $amr\left(x,\theta_t\right) \leq amr\left(x,\theta_r\right)$, the loss of x of the auxiliary model θ' should be beyond that of θ_r and θ_t, i.e., $\ell\left(x,\theta_r\right) \leq \ell\left(x,\theta_t\right) \leq \ell\left(x,\theta'\right)$. Besides, let $\alpha \overset{def}{=} -\ell\left(x,\theta_t\right) + \ell\left(x,\theta'\right)$, $\beta \overset{def}{=} \ell\left(x,\theta_r\right) - \ell\left(x,\theta'\right)$, and then $\alpha \geq 0$ and $\beta \leq 0$. Since an example's loss is at least zero and at most C, we have $-C \leq \alpha+\beta \leq 0$. Let $h(\alpha,\beta) = \frac{\sigma(\alpha)+\sigma(\beta)}{\sigma(\alpha+\beta)}$, and we have:

$$\begin{aligned} h(\alpha,\beta) &= \frac{2+e^{-\alpha}+e^{-\beta}}{1+\left(e^{-\alpha}+e^{-\beta}\right)/\left(1+e^{-(\alpha+\beta)}\right)} \\ &\leq 2+e^{-\alpha}+e^{-\beta} \\ &\leq 2+e^{C}+e^{2C}. \end{aligned}$$

Therefore, we can derive that:

$$\begin{aligned} &rmr_{\theta_r}(x,\theta_t) = \sigma(\alpha)+\sigma(\beta)-1 \\ &= h(\alpha,\beta)\,\sigma(\alpha+\beta)-1 \\ &\leq \left(2+e^C+e^{2C}\right)\sigma\left[-\ell\left(x,\theta_t\right)+\ell\left(x,\theta_r\right)\right]-1. \end{aligned}$$

□

Lemma 1. *Let θ_1 and θ_2 be different converged models under identical training settings, except that θ_1's training dataset contains a data point x, while θ_2's does not. Then, the following inequality holds:*

$$\ell(x,\theta_1) \leq \ell(x,\theta_2),$$

where $\ell(x,\theta)$ denotes the loss of model θ on the data point x.

Proof. According to the principle of Empirical Risk Minimization (ERM), a machine learning model is trained to minimize the loss over all data points in the training dataset. Formally, this can be expressed as:

$$\theta = \arg\min_\theta R(D,\theta) = \arg\min_\theta \frac{1}{N}\sum_{x'\in D}\ell(x',\theta),$$

where N represents the number of samples in D. Given the training sets D and D' for training θ_1 and θ_2, $D = D'\cup\{x\}$, the gradients of the empirical risk functions are expressed as follows:

$$\nabla_\theta R(D,\theta_1) = \frac{1}{N}\sum_{x'\in D}\nabla_\theta\ell(x',\theta_1),$$

and

$$\nabla_\theta R(D', \theta_2) = \frac{1}{N-1} \sum_{x' \in D'} \nabla_\theta \ell(x', \theta_2).$$

Given that θ_1 and θ_2 are converged, they satisfy: $\nabla_\theta R(D, \theta_1) = 0$ and $\nabla_\theta R(D', \theta_2) = 0$. under the same training conditions and assuming a smooth loss function, θ_1 can be viewed as an offset of θ_2. Specifically, we have:

$$\theta_1 = \theta_2 + \Delta\theta,$$

where $\Delta\theta$ represents the offset between the two parameter sets. Meanwhile, we have $\nabla_\theta R(D', \theta_1) = \nabla_\theta R(D', \theta_2) + H\Delta\theta$, where H denotes the Hessian matrix of $R(D', \theta)$ with respect to θ and is positive definite. Then, we obtain:

$$\Delta\theta = -\frac{1}{N-1} H^{-1} \nabla_\theta \ell(x, \theta_1).$$

By substituting $\Delta\theta$ and applying Taylor's formula, we obtain:

$$\ell(x, \theta_1) = \ell(x, \theta_2) - \frac{1}{N-1} \nabla_\theta \ell(x, \theta_2)^\top H^{-1} \nabla_\theta \ell(x, \theta_1).$$

The second-order Taylor expansion can be safely neglected, as θ_1 and θ_2 have converged and are close, making both $\Delta\theta$ and $\nabla^2_\theta \ell(x, \theta_2)$ sufficiently small. Finally, as the directions of $\nabla_\theta \ell(x, \theta_1)$ and $\nabla_\theta \ell(x, \theta_2)$ are aligned, we conclude that $\ell(x, \theta_1) \leq \ell(x, \theta_2)$.

□

Thanks to Theorem 1, we can approximate RMR using Inequality (6), and further simplify it by removing the constant coefficients and offsets. This simplification is reasonable for well-trained models which have already fit the training data with small losses. As such, we redefine RMR as:

$$rmr_{\theta_r}(x, \theta_t) \stackrel{def}{=} \sigma\left[-\ell(x, \theta_t) + \ell(x, \theta_r)\right]. \tag{7}$$

The following corollary further shows that RMR preserves the order of the absolute membership risk so that it can provide accurate comparison results of the risks of two models under membership inference attacks.

Corollary 1 (order-preserving property)**.** *Given models θ_1 and θ_2 trained on D and a reference model θ_r, for a training example x, if $amr(x, \theta_1) \geq amr(x, \theta_2)$, then $rmr_{\theta_r}(x, \theta_1) \geq rmr_{\theta_r}(x, \theta_2)$.*

Proof. Since $amr(x, \theta_1) \geq amr(x, \theta_2)$, according to Eq.(4), we have:

$$\sigma\left[-\ell(x, \theta_1) + \ell(x, \theta')\right] \geq \sigma\left[-\ell(x, \theta_2) + \ell(x, \theta')\right],$$

where θ' is an auxiliary model. By the monotonicity property of sigmoid function, we have $\ell(x, \theta_1) \leq \ell(x, \theta_2)$. So given a reference model θ_r, we have

$$\sigma\left[-\ell(x, \theta_1) + \ell(x, \theta_r)\right] \geq \sigma\left[-\ell(x, \theta_2) + \ell(x, \theta_r)\right].$$

□

Based on Eq.(7) and Corollary 1, the RMR of a target model trained on D is defined as the average risk of the whole training set given a reference model:

$$rmr_{\theta_r}(D, \theta_t) \stackrel{def}{=} \mathbb{E}_{x \sim D}\left[\sigma\left(-\ell(x, \theta_t) + \ell(x, \theta_r)\right)\right]. \tag{8}$$

The advantage of RMR for a machine learning model lies in its efficiency, as it requires only a single model to evaluate the entire training set, rather than needing an auxiliary model for each individual training example [37], [55]. Moreover, RMR strictly maintains the ranking of the absolute membership risks of target models, as established in Corollary 1, making it a lightweight relative risk indicator for membership inference attacks. As a final note, while our method derives RMR from the posterior distribution [36], this relative measure is adaptable to other attack scenarios with similar closed-form posterior probabilities as Eq.(2), which serve as the basis for Theorem 1 and Corollary 1.

IV. Reference Model Selection

In this section, we explore methods for measuring the membership risks associated with multiple target models. This scenario often arises when a user attempts training configurations, such as model architectures and hyperparameters, resulting in a set of candidate models for potential release. Beyond predictive accuracy, assessing the privacy leakage risk of these models is crucial, particularly in sensitive applications like medical or financial domains. In such scenarios, where the goal is to select the safest model against MIAs from a set of candidate models, we propose an iterative selection algorithm that identifies a reference model without requiring the retraining of an inference model.

As illustrated in Fig. 1, the candidate models are pre-trained, off-the-shelf target models provided by the user. According to Theorem 1, a valid reference model must exhibit a higher AMR, i.e., greater risk, compared to any target model as measured by RMR. A straightforward approach is to select the model with the highest RMR within the candidate set. The following corollary provides a simple method to verify whether this prerequisite is satisfied.

Corollary 2 (prerequisite of Theorem 1)**.** *Given a reference model θ_r, a target model θ_t, and a training example x, if $rmr_{\theta_r}(x, \theta_t) \leq 0.5$, then $amr(x, \theta_r) \geq amr(x, \theta_t)$. Consequently, Theorem 1 is satisfied.*

Proof. Since $rmr_{\theta_r}(x, \theta_t) \leq 0.5$, it follows that $\ell(x, \theta_r) \leq \ell(x, \theta_t)$. By the monotonicity property of the sigmoid function, for any auxiliary model θ', it holds that $amr(x, \theta_r) \geq amr(x, \theta_t)$. □

According to Corollary 2, the average risk of θ_t trained on the whole training set D is also no larger than 0.5, i.e, $rmr_{\theta_r}(D, \theta_t) \leq 0.5$. Based on this condition, we propose an iterative reference model selection algorithm that finds the riskiest model in a candidate set. In each iteration, it (1) chooses the current riskiest model as the reference model, and (2) checks whether any measurement violates the prerequisite of RMR. Since this algorithm returns the entire list of RMRs after the reference model is finalized, it also reveals the safest target model under MIAs for a user to choose for release.

The pseudo-code is presented in Algorithm 1. First, it randomly selects an initial reference model from the target model set Θ (Line 4). Then the algorithm iteratively computes the RMR of each model by (8) (Line 7). When any RMR is

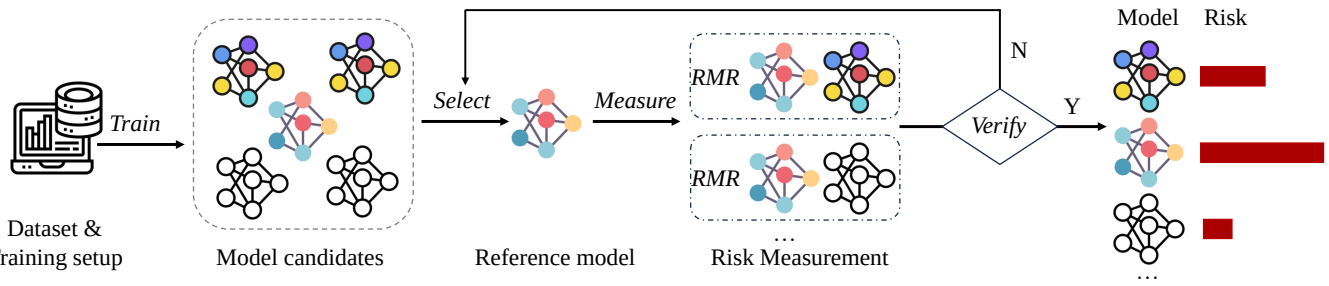


Fig. 1. The proposed algorithm provides a systematic approach for measuring membership risks across multiple target models. It operates by iteratively selecting a reference model from the set, assessing the membership risks for the target models, and verifying any violations. This iterative process continues until the algorithm identifies the reference model with the highest membership risk.

Algorithm 1: Iterative Reference Model Selection and RMR Sorting

Input: D: $\{x_1, \ldots, x_N\}$, training dataset
Θ: $\{\theta_1, \ldots, \theta_M\}$, target model set
Output: RMR measurement list R

1 $R \leftarrow \mathbf{0}_{[M]}$
2 $St \leftarrow \emptyset$ ▷ Store models sorted by RMR
3 $unserved \leftarrow \Theta$
4 $\theta_r \leftarrow randomSelect\,(\Theta)$ ▷ Initialize θ_r
5 **while** *True* **do**
6 **for** m *in* $\{1, 2, \ldots, M\}$ **do**
7 $R[m] \leftarrow rmr_{\theta_r}\,(D, \theta_m)$ ▷ Compute RMR
8 $unserved \leftarrow delete\,(unserved, \theta_r)$ ▷ θ_r served
9 $St \leftarrow sort\,(\Theta, R)$ ▷ Sort Θ by RMR
10 **for** m *in* $\{1, 2, \ldots, M\}$ **do**
11 **if** $R[m] > 0.5$ **then**
12 $\theta_r \leftarrow unserSelect\,(unserved, St)$ ▷ Fail to validate and select the riskiest model as θ_r
13 **break**
14 **if** $m == M$ **then**
15 **break** ▷ Early stop
16 **return** R

larger than 0.5, that is, unsuccessfully validated by Corollary 2, it selects a current unserved model with the highest RMR value as the reference model (Line 12) and updates the RMR list in the next iteration. Otherwise, the riskiest reference model is identified, and the algorithm will terminate (Line 15). It is easy to prove that Algorithm 1 can terminate within M (the size of Θ) iterations due to M candidate models. Therefore, considering a set of M candidate models, the time complexity of Algorithm 1 is $\mathcal{O}\left(NM^2\right)$, where $N \gg M$ is the size of the training set. Hence, when the final RMR list can be successfully validated, the algorithm can choose the riskiest model as the reference model and measure the model risk accurately. [1]

V. Empirical Evaluation

This section evaluates the effectiveness of RMR to measure the membership risk of models, and then present a case study on its application in risky example removal. We attempt to answer the following questions.

Q1: Is RMR effective in measuring membership privacy risk for a machine learning model?

Q2: How does the reference model impact RMR when evaluating a set of trained models?

Q3: What are the factors that may impact the performance of RMR when measuring a group of trained models?

Q4: How is RMR employed to pinpoint high-risk examples for an individual model?

A. Experimental Setup

Datasets. We adopt several datasets for classification tasks, namely, Purchase[2], Location[3], FMNIST, and STL10. Purchase and Location encompass tabular information, while the FMNIST and STL10 datasets consist of image samples. The Location dataset comprises check-in records, which include 4,000 data samples with 446 binary features in our experiments. The Purchase dataset contains shopping records, which consist of 39,464 data samples with 600 binary features. FMNIST is a collection of 70,000 grayscale images categorized into ten classes, with an equal number of images per class. STL10 consists of ten classes, each comprising 1,300 color images. The details are shown in Table II.

Following previous work [20], we evenly split each dataset into four equal disjoint parts: target training set, target test set, shadow training set, and shadow test set. Similar to [5], [20], [24], the number of member examples is the same as that of non-members, so a random guess results in 50% membership inference attack accuracy.

TABLE II
Dataset Statistics.

dataset	type	network	#class	#feature	#size	M
Location	Tabular	MLP	30	446	4,000	80
Purchase	Tabular	MLP	100	600	39,464	70
STL10	Image	CNN	10	27648	13,000	44
FMNIST	Image	CNN	10	1024	70,000	33

Off-the-shelf Attacks. To evaluate the effectiveness of our proposed risk measurement, we first need to determine the

[1]When the average risk of the entire training set as the validation criterion, some training examples might exhibit greater risk in a safe model than in a risk one. Nevertheless, as shown in the Section V, such exceptional cases only account for a small fraction.

[2]https://www.kaggle.com/c/acquire-valued-shoppers-challenge
[3]https://sites.google.com/site/yangdingqi/home/foursquare-dataset

ground-truth membership risk for each target model. However, since obtaining the ground-truth risk is not feasible, we adopt an alternative approach: we consider the highest attack performance as the *real membership risk* of the model, based on state-of-the-art metric-based and NN-based membership inference attacks. We consider the following six existing attacks:

- ***M-CR***: [21] relies on prediction correctness for membership inference. An input sample is inferred as a member if it is correctly predicted; otherwise, it is a non-member.
- ***M-CF***: [21], [35] infer membership by comparing the prediction confidence of a sample with a threshold. If the confidence exceeds the threshold, the sample is classified as a member; otherwise, as a non-member.
- ***M-ET***: [21] uses the entropy of confidence scores to infer membership by comparing it against a threshold.
- ***M-ME***: [21] relies on modified prediction entropy, incorporating the ground truth label and setting distinct thresholds per class. A sample is classified as a member if its modified entropy is below the threshold; otherwise, it is not.
- ***N-SM***: [5] uses a binary classifier to distinguish a target model's behavior on its training members from non-members. Following previous works [5], [20], the attack model is a 3-layer perceptron, taking the full confidence scores from the target model and the ground-truth labels as inputs.
- ***N-LiRA***: [22] adopts a statistical method to determine if a target sample was in the training set. It trains shadow models with query examples (IN) and without query examples (OUT), fits the scaled logits to two Gaussian distributions, and uses a likelihood-ratio test to infer membership. We train 16 shadow models (8 IN and 8 OUT) and evaluate the membership risk of a given model using 2,000 query examples (1,000 for the Location dataset).

In our experiments, all thresholds are optimized on the shadow datasets to achieve the best attack performance.

Training Setup. We implement different network architectures on various datasets, with the specific architecture type used for each dataset presented in Table II. The MLP is trained on tabular datasets with three hidden layers containing 1024, 512, and 256 neurons. The CNN, designed for image data, consists of three convolutional layers and two fully connected layers, with ReLU as the activation function. In the ablation study, we also evaluate our method using AlexNet and DenseNet. Each model is trained using the SGD optimizer with a learning rate of 10^{-2}, momentum of 0.9, training epoch of 100, and a batch size of 128.

Before evaluating the effectiveness of various measures, we first train M target models with different configurations, including ℓ_2-regularization, dropout, and varying network architectures. We then perform five off-the-shelf attacks to obtain the real membership risk.

Evaluation Metrics. Following existing works [5], [21], [25], [36], [37], we use Membership Accuracy (MAcc) to evaluate attack performance. MAcc is defined as the ratio of successful attacks to the total number of attacks. Given an even dataset split, we sample an example from either the training or test set with a probability of 0.5.

To evaluate the effectiveness of RMR, we use two correlation metrics: Pearson Correlation Coefficient (PCC, denoted by ρ) and Kendall Rank Correlation Coefficient (KRC, denoted by τ) [56]. PCC measures the linear relationship between two distributions, indicating how well RMR (denoted as Y) aligns with the actual membership risks (denoted as X) on the target models, expressed as:

$$\rho = \frac{\mathrm{cov}\left(\mathrm{X}, \mathrm{Y}\right)}{\sigma_X \sigma_Y},$$

where cov is the covariance, and σ_X (or σ_Y) is the standard deviation of X (or Y). And KRC measures the ordinal association between two risk ranks:

$$\tau = \frac{2}{n\left(n+1\right)} \sum_{i<j} \mathrm{sgn}\left(x_i - x_j\right) \mathrm{sgn}\left(y_i - y_j\right),$$

where $x_i \in X$, $y_i \in Y$, n is the size of X or Y, and $\mathrm{sgn}\left(\cdot\right)$ is the sign function.

Comparison Baselines. We use two risk measurement approaches as baselines. The first, referred to as *overfit* [5], [24], is a commonly employed method for evaluating the membership risk of a machine learning model. It is determined by the difference in classification accuracy between the training and test sets. The second approach, SHAPr [23], is an attack-agnostic and detailed method that approximates leave-one-out computations. For a given instance x, SHAPr extracts the feature embedding from the target model, then applies a k-NN classifier to compute x's contribution to the training set. The highest SHAPr value across all training examples is used to quantify the model's risk.

Experiment Environment. Our experiments are implemented in Pytorch and performed on an NVIDIA RTX-3090 server with Ubuntu operating system. All experiments are repeated five times and report the average results.

B. Effectiveness of RMR

To address Q1, we validate whether the relative risk measure RMR is effective in measuring the membership privacy risk of machine learning models.

Table III presents the performance of RMR and two baseline methods against real membership risks for target models. Overall, RMR consistently outperforms the baselines in both PCC and KRC metrics, with particularly strong advantanges in the image domain. For instance, on the STL10 dataset, RMR achieves a PCC of 0.88, while *overfit* scores only 0.2086, and SHAPr results in -0.0451. Fig. 2 further visualizes the distributions of attack performance and risk measures across four datasets. We observe strong correlations between the RMR measurement and the real membership accuracy, especially in the tabular datasets, where the PCCs are 0.9752 and 0.9030 for the Location and Purchase datasets, respectively. In contrast, SHAPr shows very low alignment with the real membership risk, and overfit fails to capture it in the STL10 dataset. These results highlight that RMR serves as an effective indicator of the performance of membership inference attacks.

TABLE III
RISK MEASURE PERFORMANCE COMPARISON.

Metric	Method	Location	Purchase	STL10	FMNIST
PCC	overfit	0.9455	0.8589	0.2086	0.7156
	SHAPr	-0.0144	0.1137	-0.0451	0.1840
	RMR	**0.9752**	**0.9030**	**0.8800**	**0.8609**
KRC	overfit	0.9038	0.8094	0.5378	0.7014
	SHAPr	0.4845	0.4572	0.4461	0.4779
	RMR	**0.9286**	**0.9302**	**0.8502**	**0.7102**

Fig. 2. Comparison of various risk measures with membership accuracy.

Since we use the average RMR of all training examples as the model's risk in Eq.(8), individual example risk variations may occur. In such a case, a model with higher risk might leak less membership privacy than a model deemed safer for certain examples. Table IV shows the percentage of such violations across all datasets when the riskiest model is selected as the reference model from a set of target models. Although not all training examples may strictly adhere to Corollary 2, these violations are infrequent, remaining below 10%. Furthermore, Table V presents the improvement rate of RMR when we remove these violating examples. The performance change is minimal, indicating that these violations do not significantly impact the model risk results, thereby justifying the use of the average RMR as the model risk measure.

TABLE IV
PERCENTAGE OF INDIVIDUAL EXAMPLE RISK VIOLATIONS.

	Location	Purchase	STL10	FMNIST
Viol.per	3.4%	2.6%	9.8%	3.6%

TABLE V
RMR IMPROVEMENT RATE BY REMOVING VIOLATING EXAMPLES.

Metric	Location	Purchase	STL10	FMNIST
PCC	+0.8%	+0.1%	-0.1%	-0.05%
KRC	+0.2%	+0.3%	-0.2%	+0.5%

The above results demonstrate the effectiveness of RMR in measuring the membership risk across multiple models. Additionally, it proves useful for monitoring the risk changes of an individual model throughout its training process. We prepare a reference model trained on the same training dataset without any defense mechanism and then observe the attack performance on the target model during its training process. Fig. 3 shows the change in both the RMR and the real membership risk of a single model across different epochs. We observe nearly perfect alignment between the two metrics during the training epochs, particularly in the Location and Purchase datasets. This result encourages us to use RMR as an indicator for early stopping when considering privacy leakage.

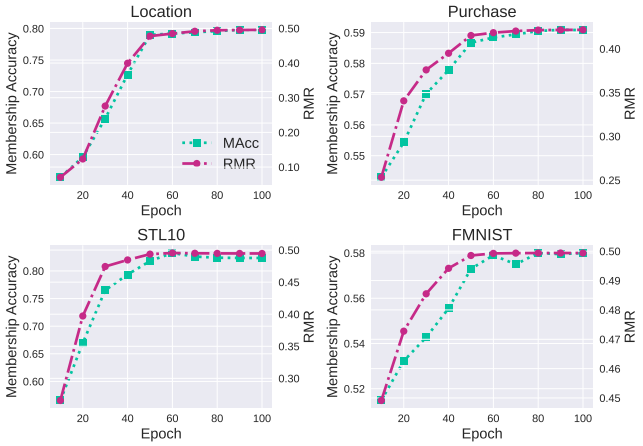


Fig. 3. RMR performance across different training epochs.

An effective membership privacy metric must also be efficient, which makes the computational cost of various measures an important consideration. To this end, we present the average time cost required to assess the membership risk of an individual model within the candidate pool in Table VI. The results indicate that RMR incurs comparable or lower costs than baselines, while offering superior performance in measuring membership risk.

TABLE VI
TIME COST COMPARISON. THE TIME COST IS STANDARDIZED BASED ON THE OVERFIT APPROACH, WITH THE VALUES IN PARENTHESES REPRESENTING THE WALL CLOCK TIME (IN SECONDS).

Method	Location	Purchase	STL10	FMNIST
overfit	1x(0.74)	1x(1.46)	1x(3.0)	1x(5.87)
SHAPr	9.5x	18.5x	8.0x	4.0x
RMR	0.5x	1.4x	0.6x	0.7x

C. Ablation Study

In this subsection, to solve Q2 and Q3, we examine possible factors contributing to the success of RMR in measuring membership risk, including the selection of the reference model, the type of network used, the training example ratio, and the number of target models.

First, we discuss the impact of the reference model on RMR. As highlighted in Section IV, the choice of reference model is crucial for the accuracy of RMR. To address this, we

TABLE VII
IMPACT OF REFERENCE MODEL SELECTION ON PCC.

	Location	Purchase	STL10	FMNIST
Iterative	**0.9752**	**0.9030**	**0.8800**	**0.8609**
Random	0.9619	0.8698	0.8636	0.7637

propose an iterative approach in Algorithm 1 for selecting the highest-risk model as the reference model. Table VII compares this iterative strategy with a random reference model selection approach in terms of PCC. Our iterative method shows a clear advantage across nearly all datasets.

TABLE VIII
RMR PERFORMANCE ON IMAGE DATASETS UNDER VARIOUS MODEL ARCHITECTURES. MIXNET INCLUDES CNN, ALEXNET AND DENSENET.

Dataset	Metric	DenseNet	AlexNet	MixNet
STL10	PCC	0.9631	0.9201	0.9617
	KRC	0.9510	0.9091	0.8896
FMNIST	PCC	0.8817	0.8924	0.8116
	KRC	0.9412	0.8030	0.7920

The previous experiments were conducted using simple architectures, such as MLP and CNN. Now, we explore the effectiveness of RMR by applying it to more complex network types and examining its performance across heterogeneous target models. We incorporate two widely adopted architectures, AlexNet and DenseNet. For each architecture type, we prepare 18 target models with varying dropout and regularization settings, and then evaluate their membership risks. Table VIII presents the performance of RMR across various network types on image datasets. On the one hand, RMR demonstrates strong effectiveness on complex models, achieving high correlation in both metrics. Furthermore, we observe that the estimation task on models with these two architectures achieves higher performance compared to a simple CNN. This discrepancy is not due to differences in network types but rather the difficulty of the evaluation task itself. Specifically, tasks are relatively easier when the differences between models are more pronounced. For instance, on the STL10 dataset, the membership accuracy variance for models with DenseNet is 0.0106, whereas for CNN, it is only 0.0004.

Furthermore, we also explore the effectiveness of RMR in scenarios where different types of network architecture are used. As shown in Table VIII, for a mixture of heterogeneous models, MixNet, RMR experiences a slight decrease in both metrics compared to the best-performing model among the three. However, it still maintains a high correlation with the real membership accuracy, demonstrating that RMR is a robust risk indicator across different architectures. Next, we discuss the impact of training examples used for RMR computation in Eq. (8). We vary the ratio of examples by randomly selecting them from the training set and evaluate the predicted membership risk using RMR. Fig. 4 shows the performance of RMR across different training example ratios. We observe that the ratio of training examples does not significantly affect the effectiveness of RMR once it exceeds 20%. However, for the small-scale STL10 dataset, fewer training examples are insufficient for accurately modeling the complex image data distribution.

Finally, we vary the number of target models and plot the performance of RMR under different target model sizes in Fig. 5. While RMR remains unaffected in simple MLP networks trained on tabular datasets, it shows improvement with more target models in CNN models trained on image datasets. This difference can be attributed to the selection of the reference model. When there is a small number of model candidates trained under similar settings, the reference model may be too close to the others in terms of membership risk, leading to more violation examples. This negatively impacts the accuracy of the estimation results. For small-scale MLPs used in Location and Purachse data, similar variations in training settings, such as dropout and ℓ_2-regularization, result in more distinct model candidates and less influenced by the number of target models. Even across different scenarios, the experimental results show that the evaluation stabilizes when the number of target models exceeds 10.

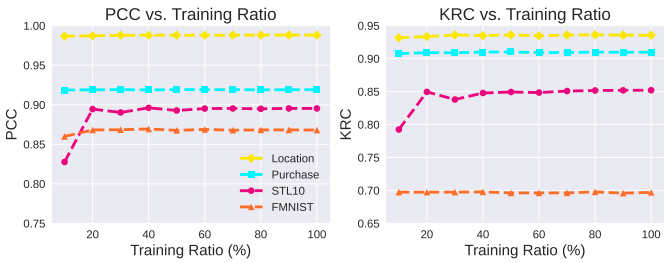


Fig. 4. Impact of training example ratio.

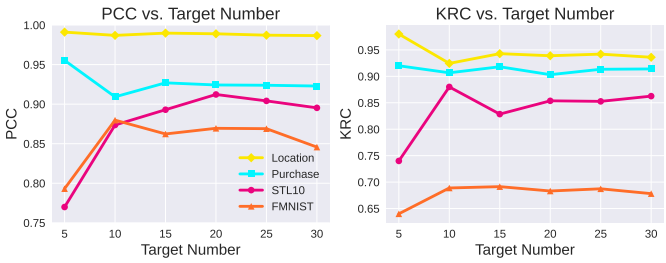


Fig. 5. Impact of target model number.

D. Case Study: Identify and Remove Risky Examples

The experimental results above demonstrate that RMR is an effective model-level risk measure. To further evaluate its effectiveness as a sample-level risk measure, we present a case study that utilizes RMR to reduce model risk by identifying and removing easy-to-attack examples in image datasets [26]. We first prepare a target model protected by defense mechanisms, such as dropout and ℓ_2-regularization, and then construct a reference model trained without any countermeasures. For this target model, we rank all training examples in descending order of their RMR values computed by Eq.(7), as determined by the predefined reference model, and subsequently remove the top-ranked examples. In addition

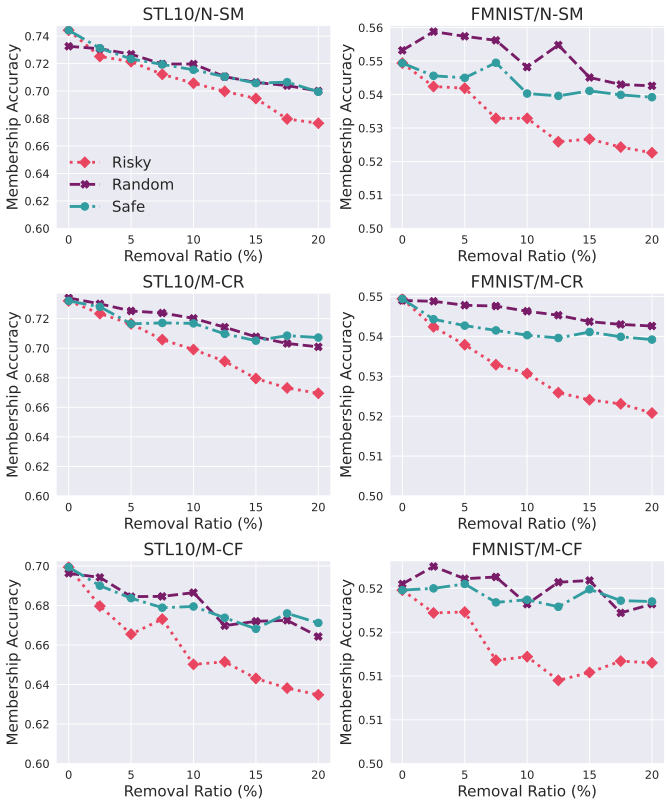


Fig. 6. Attack performance under three removal strategies.

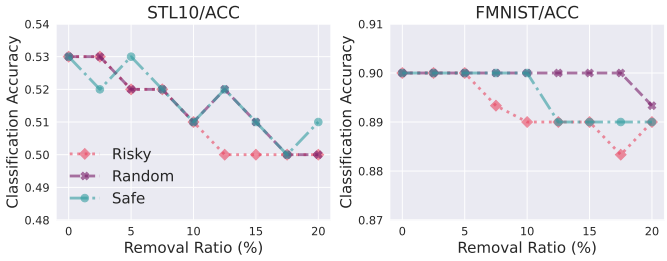


Fig. 7. Model utility under three removal strategies.

to the *risky-example-removal* strategy, we implement two additional strategies for comparison: *random-* and *safe-example-removal* (the inverse of risky-example removal). Finally, we train a new model from scratch using the modified dataset, excluding the removed examples, and evaluate its membership inference accuracy.

Fig. 6 illustrates the performance of various attacks under different removal strategies with respect to the ratio of removed examples. We observe that the risky-example-removal strategy significantly reduces the accuracy of all membership inference attacks by up to 6%, whereas the other two strategies have a smaller and comparable impact. However, the drop is still not proportional to the ratio of removed examples. For example, removing 20% of risky examples in STL10 can ideally achieve a 10% membership accuracy drop (i.e., a 100% accuracy down to a 50% random guess accuracy), but the actual drop is only 6%. There are two possible explanations for this. First, achieving 100% accuracy is not possible with current non-optimal membership inference attacks, even on the riskiest examples. Second, as recently noted by [57], the onion effect may be at play—after the riskiest examples are removed, the remaining examples (particularly those with the second-highest risk) may experience an increase in their membership risk.

Furthermore, we investigate the effect of different removal strategies on model utility and present their classification performance on the STL10 and FMNIST datasets in Fig. 7. We find that STL10 is more sensitive to the removal strategy due to its smaller training dataset compared to the FMNIST dataset. However, for both datasets, model utility is more negatively impacted when more than 10% of the risky training examples are removed. This is because examples near the decision boundary are more easily identified and are typically the most vulnerable ones [38], [39], [58]. In conclusion, there is a trade-off between membership risk and model utility when using RMR to remove risky training examples.

VI. Conclusion

Since membership inference attacks pose a fundamental threat to machine learning models trained on private data and serve as a basis for more advanced attacks, measuring the membership risk of foundation models is a critical issue. This paper presents a relative membership risk measure, RMR, that captures both record-level and model-level membership risk. Using a reference model, RMR efficiently quantifies membership risk for both training examples and machine learning models. We also propose an iterative strategy for selecting the reference model when assessing multiple target models. The experimental results demonstrate that RMR makes accurate predictions that closely align with real membership risks across various datasets and model architectures. For future work, we aim to extend RMR to address other privacy risks associated with machine learning models, such as property inference and model inversion attacks.

VII. Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No: 92270123 and 62372122), Joint Funding Special Project for Guangdong-Hong Kong Science and Technology Innovation (Grant No: 2024A0505040027), and the Research Grants Council, Hong Kong SAR, China (Grant No: 15209922, 15210023 and C2004-21GF). Meanwhile, authors thank all anonymous reviewers for their helpful suggestions to improve the paper.

References

[1] "Prediction API - pattern matching and machine learning," https://cloud.google.com/prediction/, 2016.
[2] "Amazon machine learning - predictive analytics with AWS," https://aws.amazon.com/machine-learning/, 2016.
[3] "Machine learning - Microsoft Azure," https://azure.microsoft.com/en-us/services/machine-learning/, 2016.
[4] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.

[5] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[6] H. Yan, S. Li, Y. Wang, Y. Zhang, K. Sharif, H. Hu, and Y. Li, "Membership inference attacks against deep learning models via logits distribution," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[7] A. Hu, R. Xie, Z. Lu, A. Hu, and M. Xue, "Tablegan-mca: Evaluating membership collisions of gan-synthesized tabular data releasing," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 2096–2112.

[8] J. Hayes, L. Melis, G. Danezis, and E. Cristofaro, "Logan: Evaluating information leakage of generative models using generative adversarial networks," *arXiv preprint arXiv:1705.07663*, 2017.

[9] B. Hilprecht, M. Härterich, and D. Bernau, "Monte carlo and reconstruction membership inference attacks against generative models." *Proc. Priv. Enhancing Technol.*, pp. 232–249, 2019.

[10] U. Gupta, D. Stripelis, P. K. Lam, P. Thompson, J. L. Ambite, and G. Ver Steeg, "Membership inference attacks on deep regression models for neuroimaging," in *Medical Imaging with Deep Learning*. PMLR, 2021, pp. 228–251.

[11] H. Liu, J. Jia, W. Qu, and N. Z. Gong, "Encodermi: Membership inference against pre-trained encoders in contrastive learning," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 2081–2095.

[12] Z. Zhang, M. Chen, M. Backes, Y. Shen, and Y. Zhang, "Inference attacks against graph neural networks," in *Proceedings of the 31th USENIX Security Symposium*, 2022, pp. 1–18.

[13] Y. Xiao, Q. Ye, H. Hu, H. Zheng, C. Fang, and J. Shi, "Mexmi: Pool-based active model extraction crossover membership inference," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 203–10 216, 2022.

[14] Y. Long, Z. Ying, H. Yan, R. Fang, X. Li, Y. Wang, and Z. Pan, "Membership reconstruction attack in deep neural networks," *Information Sciences*, vol. 634, pp. 27–41, 2023.

[15] H. Ko, S. Lee, Y. Park, and A. Choi, "A survey of recommendation systems: recommendation models, techniques, and application fields," *Electronics*, vol. 11, no. 1, p. 141, 2022.

[16] Z. Wang, N. Huang, F. Sun, P. Ren, Z. Chen, H. Luo, M. de Rijke, and Z. Ren, "Debiasing learning for membership inference attacks against recommender systems," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 1959–1968.

[17] M. Zhang, Z. Ren, Z. Wang, P. Ren, Z. Chen, P. Hu, and Y. Zhang, "Membership inference attacks against recommender systems," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 864–879.

[18] L. Bai, H. Hu, Q. Ye, H. Li, L. Wang, and J. Xu, "Membership inference attacks and defenses in federated learning: A survey," *ACM Computing Surveys*, vol. 57, no. 4, pp. 1–35, 2024.

[19] Y. Liu, Y. Kang, T. Zou, Y. Pu, Y. He, X. Ye, Y. Ouyang, Y.-Q. Zhang, and Q. Yang, "Vertical federated learning: Concepts, advances, and challenges," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[20] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. De Cristofaro, M. Fritz, and Y. Zhang, "Ml-doctor: Holistic risk assessment of inference attacks against machine learning models," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4525–4542.

[21] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," in *30th USENIX Security Symposium*, 2021, pp. 2615–2632.

[22] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1897–1914.

[23] V. Duddu, S. Szyller, and N. Asokan, "Shapr: An efficient and versatile membership privacy risk metric for machine learning," *arXiv preprint arXiv:2112.02230*, 2021.

[24] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 2018, pp. 268–282.

[25] Y. Long, V. Bindschaedler, and C. A. Gunter, "Towards measuring membership privacy," *arXiv preprint arXiv:1712.09136*, 2017.

[26] Y. Long, L. Wang, D. Bu, V. Bindschaedler, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "A pragmatic approach to membership inferences on machine learning models," in *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2020, pp. 521–534.

[27] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Demystifying membership inference attacks in machine learning as a service," *IEEE Transactions on Services Computing*, 2019.

[28] D. Bernau, G. Eibl, P. W. Grassal, H. Keller, and F. Kerschbaum, "Quantifying identifiability to choose and audit ϵ in differentially private deep learning," *Proceedings of the VLDB Endowment*, vol. 14, no. 13, pp. 3335–3347, 2021.

[29] B. Jayaraman, L. Wang, K. Knipmeyer, Q. Gu, and D. Evans, "Revisiting membership inference under realistic assumptions," *Proceedings on Privacy Enhancing Technologies*, 2021.

[30] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, 2017, pp. 587–601.

[31] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 267–284.

[32] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[33] F. Tramèr, R. Shokri, A. San Joaquin, H. Le, M. Jagielski, S. Hong, and N. Carlini, "Truth serum: Poisoning machine learning models to reveal their secrets," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2779–2792.

[34] Y. Chen, C. Shen, Y. Shen, C. Wang, and Y. Zhang, "Amplifying membership exposure via data poisoning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 830–29 844, 2022.

[35] A. Salem, Y. Zhang, M. Humbert, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Network and Distributed Systems Security Symposium*. Internet Society, 2019.

[36] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5558–5567.

[37] L. Watson, C. Guo, G. Cormode, and A. Sablayrolles, "On the importance of difficulty calibration in membership inference attacks," in *International Conference on Learning Representations*, 2021.

[38] Z. Li and Y. Zhang, "Membership leakage in label-only exposures," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 880–895.

[39] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *International conference on machine learning*. PMLR, 2021, pp. 1964–1974.

[40] C. Song and A. Raghunathan, "Information leakage in embedding models," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 377–390.

[41] S. Mahloujifar, H. A. Inan, M. Chase, E. Ghosh, and M. Hasegawa, "Membership inference on word embedding and beyond," *arXiv preprint arXiv:2106.11384*, 2021.

[42] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang, "When machine unlearning jeopardizes privacy," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 896–911.

[43] Q. Ye, H. Hu, X. Meng, H. Zheng, K. Huang, C. Fang, and J. Shi, "Privkvm*: Revisiting key-value statistics estimation with local differential privacy," *IEEE Transactions on Dependable and Secure Computing*, 2021.

[44] H. Zheng, H. Hu, and Z. Han, "Preserving user privacy for machine learning: Local differential privacy or federated machine learning?" *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 5–14, 2020.

[45] H. Zheng, Q. Ye, H. Hu, C. Fang, and J. Shi, "Protecting decision boundary of machine learning model with differentially private perturbation," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 3, pp. 2007–2022, 2020.

[46] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 634–646.

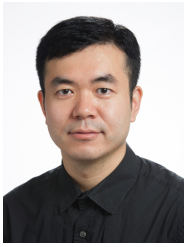
[47] V. Shejwalkar and A. Houmansadr, "Membership privacy for machine learning models through knowledge transfer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 9549–9557.

[48] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against generative models," in *Pro-*

ceedings of the 2020 ACM SIGSAC conference on computer and communications security, 2020, pp. 343–362.
[49] K. Leino and M. Fredrikson, "Stolen memories: Leveraging model memorization for calibrated {White-Box} membership inference," in *29th USENIX security symposium (USENIX Security 20)*, 2020, pp. 1605–1622.
[50] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.
[51] T. Humphries, S. Oya, L. Tulloch, M. Rafuse, I. Goldberg, U. Hengartner, and F. Kerschbaum, "Investigating membership inference attacks under data dependencies," *arXiv preprint arXiv:2010.12112*, 2020.
[52] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.
[53] M. Nasr, R. Shokri, and A. 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)*. IEEE, 2019, pp. 739–753.
[54] S. J. Russell, *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.
[55] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against generative models," in *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, 2020, pp. 343–362.
[56] H. Abdi, "The kendall rank correlation coefficient," *Encyclopedia of measurement and statistics*, vol. 2, pp. 508–510, 2007.
[57] N. Carlini, M. Jagielski, N. Papernot, A. Terzis, F. Tramer, and C. Zhang, "The privacy onion effect: Memorization is relative," *Advances in Neural Information Processing Systems*, 2023.
[58] Y. Wu, H. Qiu, S. Guo, J. Li, and T. Zhang, "You only query once: An efficient label-only membership inference attack," in *The Twelfth International Conference on Learning Representations*, 2024.

VIII. Biography Section

Li Bai is a PhD student in the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. She received the M.S. degree from the School of Cyberspace Security, University of Chinese Academy of Sciences, China, in 2019. Her research interests include data privacy and AI security, with an emphasis on membership inference attacks.



Haibo Hu is a professor with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. His research interests include cybersecurity, data privacy, and adversarial machine learning. He has published over 180 research papers in refereed journals, international conferences, and book chapters, and is granted 6 US patents and 4 China/HK patents. He is the recipient of a number of titles and awards, including IWAIT 2021 Best Paper Award, IEEE MDM 2019 Best Paper Award, WAIM Distinguished Young Lecturer, ICDE 2020 Outstanding Reviewer, VLDB 2018 Distinguished Reviewer, ACM-HK Best PhD Paper, Microsoft Imagine Cup, and GS1 Internet of Things Award. He is a senior member of ACM, IEEE and CCF, and a certified Cisco CCNA Security Trainer.

Qingqing Ye is an assistant professor in the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. She received her PhD degree from Renmin University of China in 2020. She has received several prestigious awards, including Hong Kong RGC Early Career Award, IEEE S&P Travel Award, and National Scholarship. Her research interests include data privacy and security, and adversarial machine learning.

Jianliang Xu (Senior Member, IEEE) received the BEng degree in computer science and engineering from Zhejiang University, Hangzhou, China, and a PhD degree in computer science from the Hong Kong University of Science and Technology. He is a professor at the Department of Computer Science, Hong Kong Baptist University. He held a visiting position at Pennsylvania State University and Fudan University. His research interests include data management, mobile computing, wireless sensor networks, and distributed systems.

Jin Li is currently a professor at Guangzhou University. He got his Ph.D degree in information security from Sun Yat-sen University at 2007. His research interests include design of secure protocols in Artificial Intelligence, Cloud Computing (secure cloud storage and outsourcing computation) and cryptographic protocols. He has published more than 100 papers in international conferences and journals, including IEEE INFOCOM, IEEE TIFS, IEEE TPDS, IEEE TOC and ESORICS etc. His work has been cited more than 18000 times at Google Scholar and the H-Index is 40. He is Editor-in-Chief of International Journal of Intelligent Systems. He also serves as Associate editor for several international journals, including IEEE Transactions on Dependable and Secure Computing, Information Sciences.

Chengfang Fang obtained his Ph.D. degree from National University of Singapore before joining Huawei. He has been working on security and privacy protection in several areas including machine learning, internet of things, mobile device and biometrics for more than 10 years. He has published over 20 research papers and obtained 15 patents in the domain. He is currently a principle researcher in Huawei Singapore Research Center.

Jie Shi is a Principal Researcher in Huawei Singapore Research Center. His research interests include trustworthy AI, machine learning security, data security and privacy, IoT security and applied cryptography. He has over 10 years' research experience and has published over 30 research papers in refereed journals and international conferences. He received his Ph.D degree from Huazhong University of Science and Technology, China.