Is Difficulty Calibration All We Need? Towards More Practical Membership Inference Attacks

Yu He*†
Wuhan University
Wuhan, China
yuherin@whu.edu.cn

Boheng Li*†
Wuhan University
Wuhan, China
randy.bh.li@foxmail.com

Yao Wang†
Wuhan University
Wuhan, China
valuewang@whu.edu.cn

Mengda Yang†
Wuhan University
Wuhan, China
mengday@whu.edu.cn

Juan Wang‡†
Wuhan University
Wuhan, China
jwang@whu.edu.cn

Hongxin Hu
University at Buffalo
Buffalo, NY, United States
hongxinh@buffalo.edu

Xingyu Zhao
University of Warwick
Warwickshire, UK
xingyu.zhao@warwick.ac.u

ABSTRACT

The vulnerability of machine learning models to Membership Inference Attacks (MIAs) has garnered considerable attention in recent years. These attacks determine whether a data sample belongs to the model's training set or not. Recent research has focused on reference-based attacks, which leverage difficulty calibration with independently trained reference models. While empirical studies have demonstrated its effectiveness, there is a notable gap in our understanding of the circumstances under which it succeeds or fails. In this paper, we take a further step towards a deeper understanding of the role of difficulty calibration. Our observations reveal inherent limitations in calibration methods, leading to the misclassification of non-members and suboptimal performance, particularly on high-loss samples. We further identify that these errors stem from an imperfect sampling of the potential distribution and a strong dependence of membership scores on the model parameters. By shedding light on these issues, we propose RAPID: a query-efficient and computation-efficient MIA that directly **R**e-lever**A**ges the original membersh**I**p scores to m**I**tigate the errors in **D**ifficulty calibration. Our experimental results, spanning 9 datasets and 5 model architectures, demonstrate that RAPID outperforms previous state-of-the-art attacks (e.g., LiRA and Canary offline) across different metrics while remaining computationally efficient. Our observations and analysis challenge the current de facto paradigm of difficulty calibration in high-precision inference, encouraging greater attention to the persistent risks posed by MIAs in more practical scenarios.[1]

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Machine learning**;

KEYWORDS

membership inference; difficulty calibration; computational cost

ACM Reference Format:
Yu He, Boheng Li, Yao Wang, Mengda Yang, Juan Wang, Hongxin Hu, and Xingyu Zhao. 2024. Is Difficulty Calibration All We Need? Towards More Practical Membership Inference Attacks. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3690316

1 INTRODUCTION

More personal privacy data has been incorporated into the datasets used for Machine Learning (ML) recently (such as medical [16] and communication records [8]). It is thus important to investigate whether models can effectively prevent privacy leakage. Membership Inference Attacks (MIAs) [46] have been proposed to measure the extent of a model's leakage of member samples. It aims to predict whether a given data point belongs to the training set of a given target model or not. MIAs are now the de facto standard evaluation method for models' privacy risks [37, 50] due to their simplicity to serve as a direct threat and the fact that MIAs are an important component of more sophisticated attacks [6].

Typically, MIAs exploit models' tendency to overfit their training data and therefore exhibit discrepancies in the outputs between members and non-members. Previous work seeks to learn

*Equal contribution.
†The Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University.
‡Corresponding author.

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3690316

[1]Code is available at https://github.com/T0hsakar1n/Is-Difficulty-Calibration-All-We-Need-Towards-More-Practical-Membership-Inference-Attacks.

these distinctive statistical features from the model's original outputs in different ways, such as training a binary classifier (*attack model*) [29, 38, 42, 46] or manually computing metrics like loss [63] or entropy [49]. While these attacks have demonstrated excellent performance on average-case metrics (Accuracy or ROC-AUC [44]), Carlini et al. [5] point out that they do not pose important privacy risks: high Accuracy/AUC are mainly due to the successful identification of non-members rather than members. This limitation can be attributed to the fact that the influence of samples' intrinsic difficulty [5, 41, 57] on the obtained model outputs is neglected. Specifically, certain simple non-member samples such as images with distinctive features or very short sentences, exhibit similarity to members in terms of model outputs, showing higher membership scores [17].

Difficulty calibration is proposed by Watson et al. [57] to mitigate the aforementioned issues. It attempts to quantify the difficulty of sample points (i.e., the extent to the sample represented on the whole distribution) and uses this value to regularize the model's original outputs, finally obtaining *calibrated scores* for MIAs. In practice, difficulty calibration primarily measures the difficulty of target samples by feeding them into models trained on similar data (reference models). A category of attacks known as reference-based attacks employs this technique, aiming to achieve finer-grained calibration at the cost of extensive computational resources and numerous queries [5, 41, 58, 62].

While existing reference-based attacks have achieved significant breakthroughs on recently recommended True-Positive Rate at low False-Positive Rate (*TPR at low FPR*), we argue that difficulty calibration is "not all we need" to achieve more powerful and practical MIAs. We have observed that some non-members, which could have been correctly classified, are inadvertently misclassified after difficulty calibration [57], leading to suboptimal performance. Typically, difficulty calibration assumes that outputs from reference models can effectively represent the difficulty of target samples. However, such an assumption is optimistic and thus unrealistic in many cases. In this paper, we examine and highlight that calibration errors primarily stem from two contributing factors: 1) the reference dataset is a subset sampled from the potential distribution; 2) membership scores are highly dependent on the model parameters. A more comprehensive analysis of this will be provided in Section 3.

To effectively and efficiently address this issue, we propose RAPID that directly **R**e-lever**A**ges the original membersh**iP** scores to m**I**tigate the errors in **D**ifficulty calibration, rather than treating it merely as a component of obtained calibrated scores. Specifically, while the original scores are strongly influenced by the inherent difficulty of the samples, they can provide reliable non-membership evidence because the target model directly fits member points during training. In other words, samples exhibiting extremely low original scores (e.g., high losses) are almost non-members. RAPID re-leverages the original outputs to correct misclassifications of non-members after difficulty calibration, thereby outperforming existing reference-based attacks.

To mount RAPID, we adopt a typical supervised learning approach. Concretely, the adversary first trains a surrogate model (shadow model) for the target model and several reference models. Then, the adversary evaluates the shadow dataset samples' losses (or other signal outputs) on the shadow model and reference models to obtain original membership scores and calibrated scores. The adversary can use these two scores as features to train a *scoring model*, which maps them to final scores for a threshold attack. In the end, the scoring model takes as input the target sample's calibrated scores as well as its original scores to infer the sample's membership status. The key point is that our approach introduces a **shortcut** in the inference from original outputs to membership status, allowing it to contribute independently. In contrast, previous work has been emphasizing the unreliability of original outputs and solely using them to serve as a component of calibrated scores [5, 32, 41, 57, 58, 62]. More importantly, RAPID eliminates the need for training a large number of reference models which is time-consuming, and it does not require near-unlimited query access to the target model, which is widely employed by existing state-of-the-art attacks [5, 31, 58, 62]. We leave a more detailed analysis of attack cost in Section 3 and Section 5.

We conduct extensive experiments measuring the performance of our proposed RAPID, with comparisons to other advanced attack methods. Experimental results show that RAPID achieves superior performance across various metrics while keeping its practicality in real-world scenarios. Notably, RAPID is able to achieve 5.1% TPR at 0.1% FPR on the CIFAR-10 dataset, approximately 2.5 times and 3 times the performance of state-of-the-art attacks, LiRA offline [5] and Canary offline [58]. Furthermore, RAPID shows a relative improvement of 21.4% in AUC and 14.6% in Balanced Accuracy. All improvements are achieved with only 1/25 of LiRA offline's computational cost (and potentially lower). To make stronger conclusive statements regarding the practicality of RAPID, we also evaluate it in the realm of Large Language Models (LLMs), which has seen limited exploration in prior research. Experimental results show that RAPID can achieve approximately 3 times the TPR at 0.1% FPR of attacks that only employ difficulty calibration on BERT [13]. We conduct extensive ablation studies to evaluate the influence of various components on our attack, such as the number of queries, the number of reference models, as well as the level of knowledge regarding model architecture and data distribution of the adversary. Finally, we provide additional discussion on the advantages of directly re-leveraging the original membership scores by introducing our shortcut in more complicated reference-based attacks, and fairly comparing RAPID with the most powerful (though computationally infeasible) LiRA online version. We also highlight the limitations of our work, which could provide potential directions for future research. In summary, our paper makes the following contributions:

- We discover and analyze the phenomena that inherent errors in difficulty calibration may lead to the misclassification of non-members, who could have otherwise been accurately classified, resulting in suboptimal performance.
- We propose a straightforward yet powerful MIA, known as RAPID, to address the errors in difficulty calibration, successfully outperforming other state-of-the-art attacks and being arguably more practical.
- We conduct extensive experiments in both the classic image domains and the recent field of LLMs to demonstrate the effectiveness and efficiency of our attack.

2 BACKGROUND

2.1 Machine Learning

A learned neural network in machine learning classification tasks can be represented as a parameterized function $\mathcal{M}_\theta : \mathcal{X} \to \mathcal{Y}$ that maps each input $x \in \mathcal{X}$ to a probability vector over a group of class labels $\mathcal{Y}$. θ is the set of parameters. To obtain the optimal weights θ, we utilize a dataset $\mathcal{D}$ sampled from an underlying distribution π. The process of learning the neural network model is denoted as $\mathcal{M}_\theta \leftarrow \mathcal{T}(\mathcal{D})$, where the training algorithm $\mathcal{T}$ is applied to the training set $\mathcal{D}$ to optimize the weights θ. The training process is performed by minimizing the empirical loss using the stochastic gradient descent algorithm:

$$\theta_{i+1} \leftarrow \theta_i - \epsilon \sum_{(x,y)\in\mathcal{B}} \nabla_\theta \mathcal{L}(y, \mathcal{M}_{\theta_i}(x)), \tag{1}$$

where $\mathcal{B}$ is a small batch of training samples, ϵ the learning rate for iteratively updating the parameters θ of the neural network and $\mathcal{L}$ the prediction loss such as cross-entropy loss. Utilizing the gradient descent from Equation 1 will inevitably drive the training sample's loss $\mathcal{L}(y,p)$, $x \in \mathcal{D}$ to zero. However, achieving strong generalization to unseen dataset $\mathcal{D}_{\text{test}} \in \pi$ remains a challenge in neural network training. Data augmentation [10, 52, 66] serves as an effective technique to significantly improve test accuracy. In order to make the attack scenario more realistic, we also apply this technique to the training of the target model.

2.2 Membership Inference Attacks

MIAs aim to infer whether a sample belongs to the training set of a victim model. It has drawn much attention [6, 7, 18, 20, 21, 36, 38, 48, 59, 65] because of its direct threats to privacy and its ease of deployment. Membership inference is also widely used to measure the effectiveness of machine unlearning [3] and serve as a baseline for data tracing [55, 56] and ownership verification [24, 45].

Definition of MIAs. Let π be the underlying distribution, let $\mathcal{T}$ be the training algorithm that a challenger (defender) would use, and let $\mathcal{A}$ be the attack method an attacker would use to make a prediction. The game will proceed as follows:

(1) The challenger samples a training dataset $\mathcal{D} \in \pi$ and trains a target model $\mathcal{M}_\theta \leftarrow \mathcal{T}(\mathcal{D})$.
(2) The challenger randomly flips an unbiased coin $b \in \{0, 1\}$.
(3) If $b = 0$, the challenger randomly samples a fresh target point $x \in \pi \setminus \mathcal{D}$. Otherwise, the challenger randomly samples a fresh target point $x \in \mathcal{D}$.
(4) The challenger sends (x, y) to the attacker. The attacker has black-box access to the model $\mathcal{M}_\theta$ and the distribution $\pi \setminus \mathcal{D}$.
(5) The challenger outputs a bit $\hat{b} \leftarrow \mathcal{A}^{\pi\setminus\mathcal{D}}_{\mathcal{M}_\theta}(x, y)$.
(6) if $\hat{b} = b$, output 1. Otherwise, output 0.

It is worth noting that Yeom et al. [63] also made a similar definition to ours. However, they assume the attacker has access to the entire distribution π, which means they can access potential training data before inference time. This assumption probably favors the attacker and is subject to being unrealistic. Therefore, we have modified the definition to restrict the attacker's access to only the attack dataset $\mathcal{D}_{\text{attack}} = \pi \setminus \mathcal{D}$. This modification is for assessing the attack method on unseen data. As the attack dataset $\mathcal{D}_{\text{attack}}$ and the target model $\mathcal{M}_\theta$ are typically considered fixed, the adversary's prediction can be simplified as $\mathcal{A}(x, y)$.

Attack Method. To provide a better understanding of MIAs, we make a formal definition of $\mathcal{A}$ as follows:

$$\mathcal{A}(x, y) = \mathbb{1}[\mathcal{S}(x, y) > t], \tag{2}$$

where $\mathcal{S}$ outputs the membership score of (x, y) and t indicates a threshold used for decision-making. For illustrative purposes, we start out by using the loss value [63] to represent the membership score $\mathcal{S}(x, y)$, so that $\mathcal{A}_{\text{loss}}(x, y) = \mathbb{1}[-\mathcal{L}(y, \mathcal{M}_\theta(x)) > t]$. This is intuitive as machine learning models are trained to minimize the loss of their members. Consequently, members naturally have smaller losses compared to non-members, which can be used to distinguish between them. Prior work [41, 62] has also proved that loss-threshold-based attacks are theoretically powerful. The subsequent work has built upon this and focused on high-precision membership inference [5, 31, 41, 57, 58, 62] or altering the assumptions about the attacker's knowledge (i.e., attack scenario) [9, 29, 30, 38].

Metrics. We consider the following three common metrics:

- **Balanced Accuracy.** The simplest method to evaluate attack efficacy that measures how often an attack correctly predicts membership on a balanced dataset of members and non-members [9, 18, 29, 31, 38, 41, 49, 51, 63].
- **AUC.** The most commonly used method to interpret the Receiver Operating Characteristic (ROC) curve [44] is by calculating the area under the curve (AUC). It reflects the average-case success of membership inference.
- **TPR at Low FPR.** The latest metric used to evaluate attack efficacy focuses on the TPR of the attack when the threshold t is set to a large value to achieve an extremely low FPR. This metric is currently recommended [5] because it directly reflects the actual extent of privacy leakage of the model towards its training samples.

3 RETHINKING DIFFICULTY CALIBRATION

In this section, we provide a detailed rethinking of difficulty calibration [57], which forms the basis for a variety of advanced attacks. We first provide additional background on difficulty calibration. Then, we conduct an in-depth analysis of its limitations. Finally, we present the design intuition of RAPID.

Difficulty Calibration. Watson et al. [57] have argued that $\mathcal{A}_{\text{loss}}$ is very unreliable as samples have different intrinsic difficulty. A sample exhibiting low loss is not necessarily a member; it could also be due to its low difficulty. $\mathcal{A}_{\text{loss}}$ struggle to separate these low-loss non-members from typical members since both can attain a high membership score. To improve the attack's reliability, a simple modification to the original membership score called difficulty calibration is required. It is based on the intuition that if (x, y) has low difficulty (i.e., over-represented on π), it will generally show high membership scores on all reference models trained on data similar to that of the target model. By subtracting the average of membership scores on reference models from the original membership score of the target sample, the influence of sample difficulty can be eliminated. Formally, the calibrated membership scores can

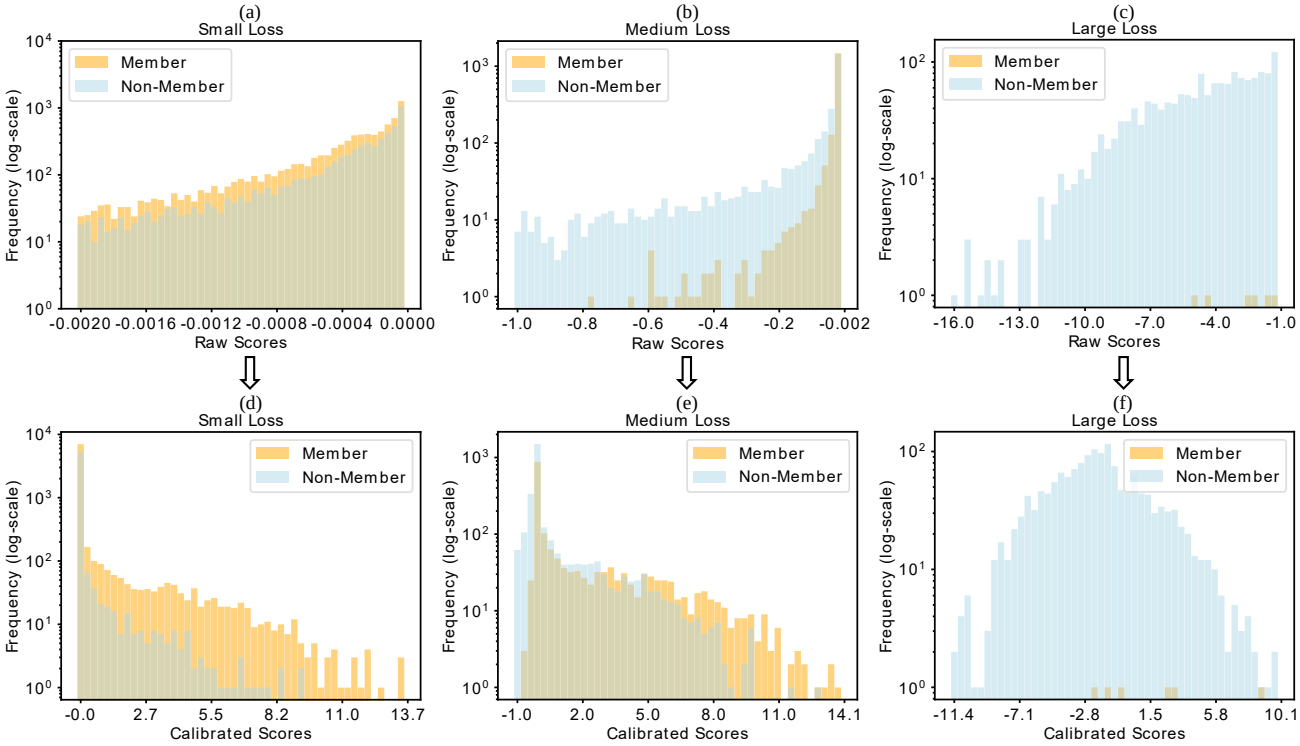


Figure 1: The distribution of raw membership scores and calibrated membership scores. All the samples with different losses obtained from the target VGG16 model are divided into three ranges: 'small loss'[0,0.002), 'medium loss'[0.002,1), and 'large loss'[1,∞). The target model is trained on the CIFAR-10 dataset. Difficulty calibration significantly increases the membership scores of some non-member samples that originally had medium or large losses.

be defined as:

$$\mathcal{S}'(x,y) = \mathcal{S}(x,y) - \mathbb{E}_{\mathcal{M}_{\text{ref}} \leftarrow \mathcal{T}(\mathcal{D}_{\text{attack}})}[\mathcal{S}(x,y)], \tag{3}$$

where the expectation is approximated by sampling several reference models from $\mathbb{T}(\mathcal{D}_{\text{attack}})$. From Equation 3 we can find that ideally, calibrated scores of non-members will approximate 0, as their original scores are mainly up to their intrinsic difficulty. Conversely, the calibrated scores of members will exceed 0 because their original scores are influenced not only by their difficulty but also by the training process itself. This technique, which we called difficulty calibration, has demonstrated breakthroughs in high-precision attacks and served as the foundation for subsequent advanced reference-based attacks [5, 41, 58, 62].

Limitations. To understand the circumstances under which the difficulty calibration succeeds or fails, we divide the samples with different losses obtained from the target model into three ranges: 'small loss'[0,0.002), 'medium loss'[0.002,1), and 'large loss'[1,∞). Specifically, Figure 1(a), Figure 1(b), and Figure 1(c) categorize samples within specific ranges of losses, while Figure 1(d), Figure 1(e), and Figure 1(f) represent the frequency distributions of calibrated scores for these samples corresponding to Figure 1(a), Figure 1(b), and Figure 1(c), respectively. To dispel potential misunderstandings, we emphasize that calibrated score distributions (i.e., the X-axis) of (d), (e), and (f) may intersect, and the different scales of the Y-axis are due to the different number of points included in them. Comparing (a) to (d), the calibrated signal indeed allows for scored highest samples to belong to the member class, making it possible to confidently predict member samples at low FPR. Therefore, if we only consider distinguishing between members/over-represented non-members, difficulty calibration indeed performs exceptionally well. However, we can observe that medium-loss and large-loss non-members, which could have been classified correctly, have a larger overlap with members in the distribution of scores after difficulty calibration. Specifically, the increased overlap area from (b) to (e) may cause degradation in metrics reflecting the average privacy leakage [57], such as Balanced Accuracy and AUC. Furthermore, the shift from (c) to (f) highlights an issue where half of the non-member samples witness a surge in their membership scores, with some even surpassing 7. High calibrated scores of large-loss non-members may render the selection of an appropriate threshold t more challenging, as the crucial metric TPR at low FPR is highly sensitive to non-members with high membership scores [5]. Overall, Figure 1 shows that depending solely on difficulty calibration constitutes a suboptimal approach, with respect to both average-case metrics and TPR at low FPR.

These limitations arise from two main factors. Firstly, the average results of membership scores obtained from reference models cannot precisely depict the extent to the target record (x, y) represented within the distribution π, as the $\mathcal{D}_{\text{attack}}$ only represents a subset of distribution π. There is an inherent difference between these two distributions. For instance, a target sample may be over-represented on a subset sampled from π but not in the entire distribution. This error can be partially mitigated by conducting multiple random samplings from $\mathcal{D}_{\text{attack}}$, provided that the attacker possesses a $\mathcal{D}_{\text{attack}}$ larger than the target model's training set. A larger $\mathcal{D}_{\text{attack}}$ implies a better approximation to the true distribution. Secondly, the calibrated scores of each sample heavily depend on the parameters of the target model and the reference models. To better illustrate this, consider the following distribution: $\mathbb{S}(x, y) = \{-\mathcal{L}(y, \mathcal{M}(x) \leftarrow \mathcal{T}(\mathcal{D})) \mid \mathcal{T} \in \mathbb{T}\}$ is the distribution of losses on (x, y) for models trained on a given dataset using different training algorithms. We follow previous work [5] to model $\mathbb{S}$ as a Gaussian distribution:

$$\mathbb{S}(x, y) \sim \mathcal{N}(\mu, \sigma^2). \tag{4}$$

For simplicity, we make an assumption that the distribution of losses on (x, y) for the target model and the reference model are independent of each other. By calculating the difference of two independent Gaussian distributions, i.e., $\mathbb{S}_{\text{tar}}(x, y)$ and $\mathbb{S}_{\text{ref}}(x, y)$, we can quantify the distribution of target samples' calibrated membership scores as:

$$\mathbb{S}_{\text{cal}}(x, y) \sim \mathcal{N}(\mu_{\text{tar}} - \mu_{\text{ref}}, \sigma^2_{\text{tar}} + \sigma^2_{\text{ref}}), \tag{5}$$

where μ_{tar}, μ_{ref}, σ^2_{tar} and σ^2_{ref} are uniquely determined by the target record (x, y) and given training sets. In a single security game, the specific parameters of the target model and the reference model actually represent a single random sampling from the distribution $\mathbb{S}_{\text{cal}}(x, y)$. Therefore, the calibrated scores depend significantly on the parameters of models and not just on membership status. For non-members, μ_{tar} and μ_{ref} should behave similarly statistically. This would make the mean of the distribution $\mathbb{S}_{\text{cal}}$ close to 0, ideally. However, the increased variance leads to a significant occurrence of calibrated scores much larger than 0, resulting in the misclassification of non-members. In worse cases, since $\mathcal{D}_{\text{attack}}$ and the target model training set do not intersect at all, it may lead to μ_{tar} being noticeably larger than μ_{ref} for non-members.

Design Intuition. Machine learning algorithms aim to minimize the loss during training, which implies that high original loss can provide sufficient evidence of non-members. On the other hand, difficulty calibration indeed helps confidently separate low-loss non-members from members. However, the aforementioned errors may cause an unexpected increase in membership scores of some high-loss non-members, making this attack suboptimal. We can thus utilize the original membership scores to differentiate between non-members who scored higher due to the difficulty of calibration and genuine members. Our method does not require training numerous reference models [5, 58, 62] to conduct a parametric likelihood-ratio test or querying the target model extensively [31, 62] to mitigate the influence of target model parameters. By introducing a **shortcut** in the inference from original outputs to membership status, our proposed RAPID can significantly (as shown in our experiments) enhance the performance of MIAs while ensuring their practicality.

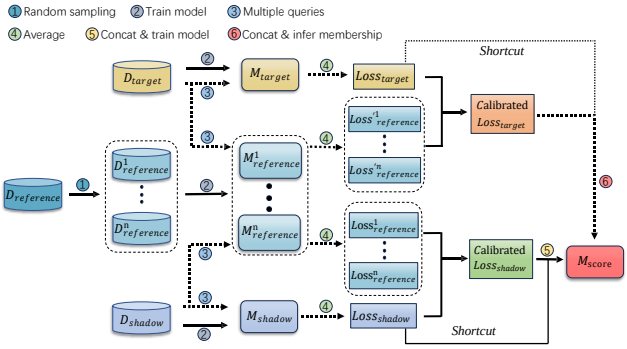


Figure 2: General attack pipeline of our RAPID.

4 ATTACK METHODOLOGY

In this section, we present the methodology of RAPID. We begin by defining the threat model that our attack operates under. Then we outline the pipeline of RAPID when the attacker has only black-box access to the target model. Finally, we introduce some useful techniques for enhancing the attack performance.

4.1 Threat Model

In this paper, we primarily focus on the most commonly adopted black-box setting, in which the attacker only has access to the posterior probability distribution of the target model's outputs. We also follow previous advanced works [5, 29, 31, 38, 41, 42, 46, 57, 58, 62], assuming that the attacker can sample sufficient data from $\pi \setminus \mathcal{D}$ and knows the target model's architecture. We will show these two assumptions can be relaxed in Section 5. Recently, there have been extensions of MIAs from black-box scenarios to settings such as white-box scenarios [29, 38] and label-only [9, 30], which will not be discussed in this paper.

4.2 Attack Method

We suggest training a scoring model $\mathcal{M}_{\text{score}}$ to map the original membership scores $\mathcal{S}(x, y)$ and the calibrated membership scores $\mathcal{S}'(x, y)$ together to the final membership scores, which are then used for membership inference. Therefore, the definition in Equation (2) can be modified as follows:

$$\mathcal{A}(x, y) = \mathbb{1}\left[\mathcal{M}_{\text{score}}(\mathcal{S}, \mathcal{S}') > t\right]. \tag{6}$$

We train a model $\mathcal{M}_{\text{score}}$ to find the optimal mapping toward final membership scores because original membership scores and calibrated membership scores have different scales. The scoring model learns to directly correct the prediction errors caused by the aforementioned errors in $\mathcal{S}'(x, y)$ using $\mathcal{S}(x, y)$. Specifically, instances with high calibrated membership scores but low original membership scores are expected to be non-members rather than members. Using a heuristic search algorithm to obtain the parameters required for optimal mapping is apparently suitable in this case. To conduct our attack, the adversary needs to perform four steps: shadow model training, reference model training, scoring

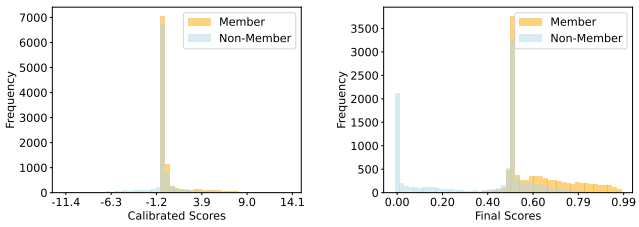


Figure 3: The frequency distributions of final scores and calibrated scores, which were sampled using a VGG16 model trained on CIFAR-10.

model training, and membership inference. We give the detailed pipeline of our proposed RAPID in Figure 2.

Shadow Model Training. As the attacker does not have access to the target model's training dataset $\mathcal{D}_{\text{target}}$. We thus sample a subset $\mathcal{D}_{\text{shadow}}$ from $\mathcal{D}_{\text{target}}$'s i.i.d. (independent identically distributed) dataset $\mathcal{D}_{\text{attack}}$ to train the shadow model $\mathcal{M}_{\text{shadow}}$. It shares similarities with the target model $\mathcal{M}_{\text{target}}$ in terms of properties, and we can utilize its outputs and $\mathcal{D}_{\text{shadow}}$ to train $\mathcal{M}_{\text{score}}$.

Reference Model Training. The attacker then uses another subset of $\mathcal{D}_{\text{attack}}$, referred to as $\mathcal{D}_{\text{reference}}$ to train reference models $\mathcal{M}_{\text{reference}}$. Different training algorithms $\mathcal{T}$ can be used to obtain reference models with different parameters. See more detailed discussion in Section 4.3.

Scoring Model Training. The attacker trains a scoring model, namely $\mathcal{M}_{\text{score}}$, using the attack dataset $\mathcal{D}_{\text{attack}}$. The scoring model is modeled as a Multi-Layer Perceptron (MLP) with a single output channel. To confine the output within the range of [0,1], a sigmoid layer is applied to the model's output. $\mathcal{M}_{\text{score}}$ is thus defined as follows:

$$\mathcal{M}_{\text{score}} = sigmoid(\mathcal{MLP}(\mathcal{S} \oplus \mathcal{S}')). \tag{7}$$

The model takes as input the concatenation of the original membership scores and the calibrated membership scores of the samples, while the corresponding labels are set to 1 if the sample belongs to the training set of the shadow model, and 0 otherwise. Binary Cross-Entropy Loss is utilized to compute the loss, and the objective is to minimize $\mathcal{L}(\mathcal{M}_{\text{score}}(\mathcal{S}, \mathcal{S}'), label)$ during training. Figure 3 demonstrates that the final membership scores obtained from $\mathcal{M}_{\text{score}}$ exhibit a higher level of discrimination between members and non-members compared to the calibrated membership scores. It is worth noting that Yuan et al. [64] proposed a self-attention-based attack that utilizes the transformer [53] to capture global dependencies among inputs and enables interaction within the inputs. We also experiment with modeling the scoring model as a transformer, but the results show no significant improvement. We believe this is because mapping the original scores and the calibrated scores to the optimal final scores is a simple task that an MLP can perform well. Future work will be conducted to further investigate the specific impact of the scoring model's architecture on attack performance.

Performing Membership Inference. The attacker is finally able to conduct MIAs on each given sample. By feeding the target sample to both $\mathcal{M}_{\text{target}}$ and $\mathcal{M}_{\text{reference}}$, the attacker obtains $\mathcal{S}(x, y)$ and $\mathcal{S}'(x, y)$ respectively. These scores are concatenated and then input into $\mathcal{M}_{\text{score}}$ to obtain the final membership score. To achieve optimal attack accuracy, the attacker simply needs to set the threshold t to 0.5. Through sweeping over a range of values for the threshold t, the adversary can obtain the tradeoff between FPR and TPR, allowing for the calculation of AUC and the attack's TPR at a given low FPR. Compared to prior works, our attack method has an additional advantage: previous reference-based attacks do not provide guidance on the appropriate threshold t to achieve the desired low FPR attack in real scenarios [5, 33, 41, 57, 58]. However, in our method, the shadow model can be utilized to determine an appropriate t in order to achieve the target FPR.

4.3 Generic Techniques

We will introduce two techniques used in our complete attack that significantly enhance the attack performance. We argue that these techniques could potentially serve as generic building blocks to enhance a variety of reference-based attacks.

Random Sampling. As the size of $\mathcal{D}_{\text{attack}}$ increases, the extent to the target record represented in the distribution π can be approximated better using $\mathbb{E}_{\mathcal{M}_{\text{ref}} \leftarrow \mathcal{T}(\mathcal{D}_{\text{attack}})}[\mathcal{S}(x, y)]$. Specifically, when the adversary possesses $\mathcal{D}_{\text{attack}}$ significantly larger than $\mathcal{D}_{\text{target}}$, the attack can be improved by training several $\mathcal{M}_{\text{reference}}$ on different subsets of $\mathcal{D}_{\text{attack}}$ sampled each time randomly. In more realistic scenarios where the attacker's available data is insufficient for random sampling, a substitute approach to improve the attack performance is by training multiple reference models using different training algorithms $\mathcal{T}$, such as varying initialization parameters. This works as it brings the computed inherent difficulty (i.e., the average outputs) of the target record closer to μ_{ref}, thereby partially mitigating the impact resulting from the randomness in sampling from a distribution $\mathbb{S}_{\text{ref}}$.

Multiple Queries. To make the observed original membership score of the target record closer to its μ_{tar}, a naive idea is to train multiple $\mathcal{M}_{\text{target}}$ on $\mathcal{D}_{\text{target}}$ and performs a single query on each of them. However, this appears to be an unfeasible strategy for the adversary. We can thus alternatively enhance the attack by averaging the outputs obtained from multiple queries on the same model on the target record and its augmentations. Notably, Carlini et al. [5] also apply this method to fit multiple-dimensional spherical Gaussians. They argue that these perturbed inputs may be seen by the target model during training and thus contain additional membership signals. However, we observe that simply averaging the outputs can already greatly enhance the attack, which is not entirely consistent with the explanation provided by Carlini et al.. We believe that this enhancement comes from mitigating the errors caused by the dependence of the target point's membership score on the parameters θ of the target model. We leave a detailed discussion in Section 6.3. As the method of fitting multiple-dimensional spherical Gaussians applies only to attacks using Gaussian likelihood estimate [5, 58], we argue that subsequent MIAs can use the average membership scores obtained from multiple queries to replace the original scores to enhance their performance.

Table 1: The prediction accuracy of different model architectures on different datasets.

Dataset	CIFAR-10		CIFAR-100		CINIC-10		SVHN	
Model	Train acc	Test acc	Train acc	Test acc	Train acc	Test acc	Train acc	Test acc
MobileNetV2	99.8%	84.1%	100.0%	55.1%	94.5%	79.7%	100.0%	95.2%
VGG16	99.8%	82.4%	99.9%	48.5%	99.9%	80.0%	100.0%	94.7%
ResNet50	98.1%	75.0%	100.0%	41.1%	99.8%	79.8%	99.8%	94.2%
DenseNet121	100.0%	81.6%	100.0%	44.9%	100.0%	80.2%	100.0%	94.9%

5 EVALUATION

In this section, we evaluate our RAPID on various benchmark datasets and diverse model architectures. We focus on three standard metrics: Balanced Acc, AUC, and TPR at low FPR, which have been detailed in Section 2. Through extensive experiments, we demonstrate that our attack outperforms the state-of-the-art methods and has lower attack costs. We also evaluate our attack under the defense of Differential Privacy (DP), which is a widely applied defense mechanism against privacy leakage attacks. In addition to the standard evaluation work emphasized by existing MIAs, we also conduct evaluations in the field of LLMs to explore the practicality of our attack. To distinguish these results from the classic evaluations, the specific experimental setup and results for this section are presented in Section 5.3.

5.1 Experimental Setup

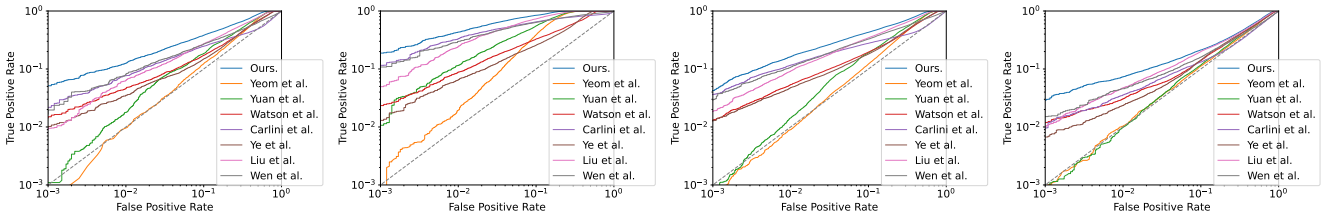
Datasets. In the main experimental section, we select four benchmark image datasets, namely CIFAR-10 [26] (a benchmark dataset used for classification tasks), CINIC-10 [11] (an extension of CIFAR-10 consisting of 270,000 images, with downsampled ImageNet images for the same classes), CIFAR-100 [26] (similar to CIFAR-10 but with 100 classes), and SVHN [39] (consisting of 99,289 color images of house numbers from the Google Street View dataset). Additionally, we also choose two text datasets, which are used for training classification models and testing attacks, including Location [60, 61] (containing location "check-in" records of mobile users in the Foursquare social network) and Texas [1] (presented in the Hospital Discharge Data Public Use Data File provided by the Texas Department of State Health Services). All datasets are divided into three equal-sized parts: $\mathcal{D}_{\text{target}}$, $\mathcal{D}_{\text{shadow}}$, and $\mathcal{D}_{\text{reference}}$. We have observed that some previous work may overlook the significant impact of the size of the attack dataset on the attack performance [31]. Generally, a larger attack dataset leads to a better approximation of sample difficulty. Therefore, when evaluating various attacks, it is crucial to ensure that different attack methods have seen an equal number of samples during the training process.

Network Architecture. For image datasets, we consider four commons architectures: VGG16 [47], ResNet50 [19], DenseNet121 [23], and MobileNetV2 [43]. For text datasets, we train a model with two fully connected layers for classification. We use the SGD algorithm to train the models, with a learning rate (lr) set to 0.1, momentum set to 0.9, and weight decay [27] set to 5e-4. We also apply a cosine learning rate schedule [34] for optimization. Data augmentation [10] is enabled during the training of the target models to enhance their generalization. For the scoring model, we train a 4-layer MLP with a single output channel.

Attack Baselines. We compare our RAPID with seven state-of-the-art or representative attack methods. Among them, Yeom et al. [63] leverage the loss of the target model for decision boundary estimation. Yuan et al. [64] propose a new signal called sensitivity, which exhibits a larger gap between members and non-members. Both Watson et al. [57] and Ye et al. [62] employ difficulty calibration. The latter emphasizes the reliance of membership scores on the target model and uses models distilled from the target model as reference models. Liu et al. [31] take the first step to exploit the information from the training trajectory to conduct membership inference attacks and achieve advanced performance. Both the methods proposed by Ye et al. and Liu et al. rely on the property that the self-distilled reference model is similar to the target model so that for non-members μ_{tar} and μ_{ref} will be closer. We also compare RAPID with LiRA proposed by Carlini et al. [5] and Canary proposed by Wen et al. [58], both using Gaussian likelihood estimate and currently achieving the best performance. Specifically, LiRA utilizes augmentations of the target sample to compute statistics, while Canary uses adversarial tools to directly optimize for queries that are discriminative.

Attack Setup. In the main experiments, we train 4 reference models on $\mathcal{D}_{\text{reference}}$, each with a different random initialization. We compute the average of the membership scores obtained from these reference models to calculate the calibrated membership scores. For all models, we utilize the multiple queries technique, where the average membership scores obtained from all queries represent per model's output scores. To ensure fairness in comparison, we re-implement existing attacks (except LiRA and Canary) using the same number of reference models as ours, if they have claimed that their performance is related to the number of reference models in the papers. When evaluating LiRA and Canary, we follow the original papers' setting for the number of reference models (128 and 64 respectively). We also ensure that LiRA queries the same number of times as RAPID (8 times). Since optimizing hyperparameters significantly affects the performance of Canary, we follow its original settings.

Why Not Online Version? In practice, both LiRA and Canary have an offline version and an online version. For instance, LiRA online firstly trains 256 reference models, half of which are *IN models* trained on datasets that include the target record, and the other half are *OUT models* trained on datasets that do not include the target record. Then, LiRA fits two Gaussian distributions to the confidences of the IN and OUT models on the target record. Finally,



(a) CIFAR-10 (b) CIFAR-100 (c) CINIC-10 (d) SVHN

Figure 4: The ROC curves of attack results on VGG16 models trained on four benchmark datasets.

Table 2: The attack performances of different attacks on VGG16 models trained on four benchmark datasets. Additional attack results for other model architectures can be found in Appendix.B of our technical report [22].

Attack	TPR @ 0.1% FPR				AUC				Balanced Accuracy			
method	CIFAR-10	CIFAR-100	CINIC-10	SVHN	CIFAR-10	CIFAR-100	CINIC-10	SVHN	CIFAR-10	CIFAR-100	CINIC-10	SVHN
Yeom et al. [63]	0.0%	0.1%	0.1%	0.1%	0.643	0.866	0.660	0.552	64.1%	83.4%	65.8%	55.8%
Yuan et al. [64]	0.1%	1.0%	0.1%	0.1%	0.680	0.895	0.691	0.562	64.6%	84.0%	66.1%	55.7%
Watson et al. [57]	1.4%	2.0%	1.3%	1.2%	0.629	0.750	0.645	0.566	58.6%	68.9%	58.9%	53.4%
Carlini et al. [5]	2.2%	11.5%	3.6%	1.0%	0.534	0.807	0.547	0.500	57.5%	78.0%	59.0%	52.1%
Ye et al. [62]	1.0%	1.2%	1.3%	0.5%	0.629	0.752	0.649	0.571	59.2%	71.9%	59.3%	53.3%
Liu et al. [31]	0.9%	4.2%	1.8%	1.0%	0.708	0.929	0.755	0.600	64.2%	85.4%	67.2%	56.0%
Wen et al. [58]	1.7%	9.1%	2.9%	1.5%	0.610	0.837	0.665	0.532	59.0%	77.6%	62.9%	53.5%
Ours	**5.1%**	**18.8%**	**4.9%**	**2.9%**	**0.776**	**0.958**	**0.799**	**0.618**	**69.1%**	**89.1%**	**70.5%**	**57.1%**

Table 3: Time cost of all attacks against a VGG16 model trained on CIFAR-10.

Attack Method	[63]	[64]	[57]	LiRA offline [5]	LiRA online [5]	[62]	[31]	Canary offline [58]	Canary online [58]	**ours.**
Time Cost/h	0.22	0.47	0.46	13.82	>200000	0.87	0.51	38.5	>100000	**0.58**

it queries the confidence of $\mathcal{M}_{\text{target}}$ on the target record and outputs a likelihood-ratio test. However, LiRA offline only trains 128 OUT models and outputs a one-sided hypothesis test. When evaluating LiRA and Canary in our main experiments, we implement the offline version of them. This is because LiRA (Canary) online requires training 128 (64) IN models for each target sample, which is not so feasible for common attackers. More discussions on why comparing with LiRA (Canary) offline is reasonable can be found in Attack Cost Analysis of Section 5.2.

5.2 Experimental Results

Finally, we present the performance of our attack in the black-box scenario, comparing it to the seven advanced baselines [5, 31, 57, 58, 62–64]. Furthermore, we provide a detailed attack cost analysis of all attacks. Lastly, we provide the results of all attacks against models using DP-SGD [2]. Table 1 reports the accuracy of $\mathcal{M}_{\text{target}}$.

Main Evaluation. Compared to the latest representative works, our proposed attack outperforms. Figure 4 demonstrates the superior performance of our attack in the low FPR regime. This holds even versus the attacks using Gaussian likelihood estimate [5, 58], which require training a large number of reference models. Table 2 presents the same advanced performance of our attack in terms of average metrics, surpassing previous attacks [5, 31, 57, 58, 62–64] by a significant margin. For example, over CIFAR-100 RAPID elevates the best TPR @0.1% FPR from 11.5% to 18.8%, best AUC from 0.929 to 0.958, and best Acc from 85.4% to 89.1%. We posit that this represents a significant advancement in the MIA domain, as previous work has struggled to achieve optimal performance across all metrics simultaneously. For instance, while LiRA and Canary offline achieve significant breakthroughs in currently recommended TPR at 0.1% FPR, they even fall short of the initial loss attack [63] in metrics reflecting average-case success. Additional attack results for other model architectures and datasets can be found in Appendix.B and Appendix.C of our technical report [22].

Attack Cost Analysis. To shed light on the practicality of existing state-of-the-art work [5, 31, 58, 62] and our proposed attack, we will provide an analysis in two aspects: query cost and computational cost. In distilled-based attacks [31, 62], assuming the distillation dataset size is N and the number of distillation rounds is E, if the attack targets n records, the attack would require $NE+n$ queries on the target model. In contrast, our attack only requires $8n$ queries on the target model. Specifically, taking CINIC-10 as an example, the cost of our attack for 20000 sample points reduces to approximately 1/42. As for computational cost, our proposed RAPID only requires training 4 reference models and a shadow model to achieve better

Table 4: Attack performance of RAPID against a DenseNet121 model trained on CIFAR-10 using DP-SGD.

Noise Multiplier (σ)	ϵ	C = 10 Model Acc	Attack Acc	TPR @ 0.1% FPR	Attack AUC
0.0	∞	77.1%	67.7%	3.0%	0.756
0.1	>5000	66.8%	54.0%	0.2%	0.552
0.2	>1000	58.6%	51.2%	0.2%	0.516
0.5	>100	44.9%	50.4%	0.2%	0.507
1.0	8	30.2%	49.9%	0.1%	0.502

performance compared to LiRA [5], which requires training at least 128 models. This reduces the computational cost to approximately 1/25 (and potentially lower). To provide an explicit time complexity analysis for all attacks, we report the total time cost of various attacks against a VGG16 model trained on CIFAR-10 using a single NVIDIA GeForce RTX 3070 Ti in Table 3. The time cost of LiRA (Canary) online is approximately proportional to the number of samples attacked. Thus, its time cost is calculated theoretically by measuring the time required to attack one sample. The astronomical computational overhead of LiRA (Canary) online renders it an infeasible attack—the adversary needs to train 128 IN models for each potential member at inference time. Therefore, we use LiRA (Canary) offline as the state-of-the-art baselines in our main experiments.

In practice, Carlini et al. [5] use a clever method that circumvents the necessity to train 128 models for each point to evaluate the theoretical performance of LiRA online. However, we note that the implementation in their code repository[2] 1) relaxes the assumption that the attack set contains only non-members, and (2) potentially makes LiRA online advantaged more (than traditional implementations) as the IN/OUT models are highly similar to the target model. To ensure fairness in comparison, we have placed a detailed discussion of the theoretical performance gap between RAPID and LiRA online in Section 8.

Attack Against DP-SGD. Differential privacy [15] is a widely used defense mechanism against all privacy leakage attacks [25, 28, 48]. It imposes theoretical bounds on the success rate of MIAs by directly restricting the ability to distinguish between two neighboring datasets (differing only in the inclusion or exclusion of a particular sample). This is directly related to MIAs. Previous studies have also explored this scenario [5, 31], and we follow their investigation to examine the defensive effect of the DP-SGD training algorithm [2] on our attack. We fixed the clipping norm to 10 and evaluated the performance of prior works and our attack on a DenseNet121 model trained on the CIFAR-10 dataset. The privacy budget ϵ can be controlled by varying the noise multiplier parameter. From Table 4 and Figure 5, we can observe that DP-SGD indeed effectively defends against all MIAs. However, DP-SGD significantly reduces the classification accuracy of the target model under high clipping norms, even when the noise multiplier is set to 0.1. We should thus carefully consider the trade-off between the defense level achieved by differential privacy and the loss of model accuracy. We primarily focus on the scenario where σ (noise multiplier) is set to 0.1 to evaluate the defense level of DP against existing attacks since this

[2]https://github.com/tensorflow/privacy/tree/master/research/mi_lira_2021

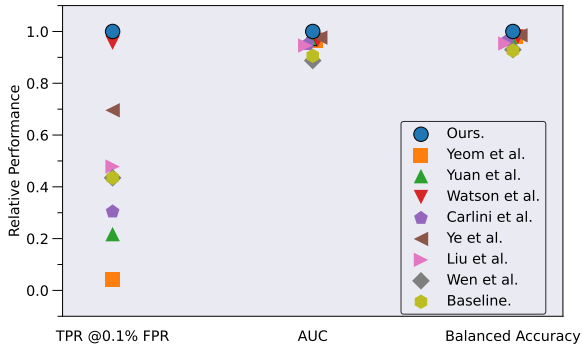


Figure 5: Attack performance of prior works and our attack against a DenseNet121 model trained on CIFAR-10 using DP-SGD. The noise multiplier σ is set to 0.1. Additional attack results for other σ can be found in Appendix.D of our technical report [22].

setting maintains an acceptable model accuracy. It can be observed that while the gap between different attacks has narrowed, our attack continues to outperform other works across all metrics. Our work presents a greater challenge to DP-SGD in the better trade-off between defense level and model performance.

5.3 Attack Against LLMs

In the realm of LLMs, MIAs can assess the degree of privacy leakage in both the pre-training and fine-tuning stages. Pre-training is primarily conducted on publicly available datasets, and the data used to train the model is often public knowledge. Fine-tuning typically occurs on smaller, and more private datasets. Therefore, our primary focus is on the fine-tuning phase, where full model fine-tuning and prompt-based learning [4, 40] are two commonly used methods. Previous work has already pointed out that the privacy risk of prompted models exceeds that of fine-tuned models at the same utility levels [14], and we are thus interested in whether our proposed RAPID can launch an effective attack against fine-tuned LLMs or not.

Setup. We fine-tune BERT [13] to solve three standard downstream text classification tasks: cola [54], cb [12], and mrpc [54]. This is because BERT has demonstrated strong generalization capabilities on these classification tasks. To investigate the impact of model size on membership risk, we conduct attack evaluations on both the BERT-base version (total Parameters=110M) and the BERT-large version (total Parameters=340M). Within all our experiments, the learning rate (lr) is set to 3e-5 and weight decay [27] is set to 5e-4 in the training process. We fine-tune the model for 20 epochs and use the checkpoint with the highest validation accuracy during tuning. We report the fine-tuning results in Table 5. In the attack setup, we follow the data splitting method outlined in Section 5.1 and only fine-tune two reference models on $\mathcal{D}_{\text{reference}}$ for our attack. The technique of multiple queries is not employed because there is no natural data augmentation available in the text domain as there is in the image domain. However, Mattern et al. [35] have

Table 5: The classification accuracy of BERT-base and BERT-large fine-tuned on different datasets.

Dataset	cola		cb		mrpc	
Model	Train acc	Test acc	Train acc	Test acc	Train acc	Test acc
BERT-base	98.0%	80.4%	99.8%	79.2%	99.5%	77.8%
BERT-large	100.0%	86.3%	99.9%	82.6%	99.7%	81.4%

Table 6: The attack results of BERT-base and BERT-large fine-tuned on mrpc.

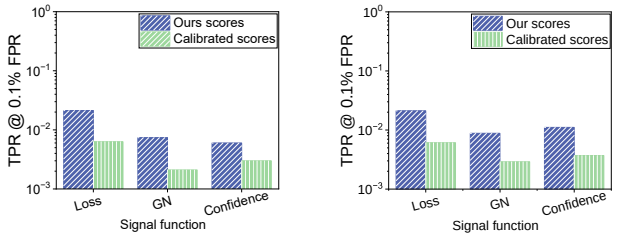
	TPR @ 0.1% FPR		AUC		Balanced Accuracy	
Attack Method	BERT-base	BERT-large	BERT-base	BERT-large	BERT-base	BERT-large
Duan et al. [14]	0.2%	0.1%	0.686	0.689	63.1%	59.9%
Watson et al. [57]	0.4%	**0.2%**	0.654	0.654	59.0%	58.4%
Ours	**1.1%**	**0.2%**	**0.745**	**0.700**	**66.7%**	**60.1%**

recently proposed a neighborhood attack that uses synthetically generated neighboring texts. This aligns closely with our idea, implying RAPID's potential for further enhancement in attacking LLMs.

Experimental Results. We compare our attack to the original loss-based attack in [14] and attacks with difficulty calibration in [35, 57] as other baselines do not take this scenario into consideration. The results in Table 6 demonstrate that RAPID still outperforms other baselines in attacking well-fine-tuned LLMs. However, the advantage of our attack is observably reduced compared to that of the computer vision domain, especially in terms of TPR at low FPR. One possible reason is that LLMs, due to their strong generalization capabilities obtained from the pre-training phase, result in small prediction losses for most non-members. In other words, the number of misclassified non-members (due to difficulty calibration) that can be directly corrected using the original membership scores is smaller. This is consistent with the worse TPR results for BERT-large compared to BERT-base, as BERT-large has a larger model capacity and stronger generalization abilities. Note that Carlini et al. [6] have demonstrated that larger pre-trained language models would memorize more training data, which contrasts with the experimental results in Table 6. We speculate that this is because the memorization principles of LLMs differ during the pre-training and fine-tuning stages. We have also observed that even with a larger training-testing accuracy gap compared to models trained on SVHN (see Table 1), the TPRs of all attacks against LLMs become generally worse, which contradicts traditional views. We hypothesize that the target dataset itself probably has a quite small proportion of outliers (hard samples), making the distribution of outputs for member points similar between the target models and reference models. We argue that **the inherent distribution properties of the dataset also significantly influence the attack's TPR at a given FPR, not only the level of overfitting.** For more experimental results, please refer to Appendix.E of our technical report [22].

6 ABLATION STUDY

In this section, we conduct extensive experiments to investigate the specific impact of each component on the final performance. We aim to further substantiate our explanation in Section 3 regarding the suboptimality of difficulty calibration. Specifically, we start by exploring the impact of reference models and the signal function employed. Then we discuss the effects of random sampling and the number of $\mathcal{M}_{\text{reference}}$ on the attack. We also examine the impact of varying numbers of queries. Lastly, we attempt to relax two common assumptions regarding the same architecture to $\mathcal{M}_{\text{target}}$ and i.i.d. $\mathcal{D}_{\text{attack}}$ used by the attacker to demonstrate the efficacy of our attack in more realistic scenarios. In our ablation studies, we utilize the CINIC-10 dataset by default unless otherwise stated.



(a) Trivial Model **(b) Distilled Model**

Figure 6: Our proposed method significantly improves the TPR at low FPR compared to solely using calibrated membership scores.

6.1 Reference Model and Signal Function

In practice, the adversary can select different reference models and signal functions for difficulty calibration. Common reference models include models trained from scratch on $\mathcal{D}_{\text{reference}}$ (i.e., trivial models [57]) and models distilled from the target model (i.e., distilled models [62]). As for signal functions, common options include loss [63], confidence [42], and gradnorm [38]. To investigate whether intrinsic errors in difficulty calibration are a pervasive phenomenon, we compare the different performances of attacks using only calibrated scores and attacks re-leveraging original scores across various reference models and signal function settings. Note that to emphasize the direct impact of the shortcut introduced by us, other enhancement techniques such as random sampling and multiple queries are not utilized. Table 7 and Figure 6 demonstrate that introducing a shortcut of $S_{target}(x, y)$ effectively enhances the performance across all evaluation metrics. This justifies our claim that difficulty calibration represents a suboptimal approach and that original membership scores can directly correct errors it generates.

6.2 Random Sampling

We have argued the adversary can significantly enhance the attack by training several $\mathcal{M}_{\text{reference}}$ using random sampling when he has a larger $\mathcal{D}_{\text{attack}}$ compared to $\mathcal{D}_{\text{target}}$. In the worst case, the attacker can also achieve a slightly weaker improvement by changing the initialization parameters of these $\mathcal{M}_{\text{reference}}$ trained on $\mathcal{D}_{\text{reference}}$. We are interested in understanding the impact of the number of $\mathcal{M}_{\text{reference}}$ on the attack results, both with and without random sampling. Figure 7 illustrates the TPR of our attack at a fixed FPR

Table 7: Comparison of the performance using prior calibrated membership scores and ours on a VGG16 model trained on CINIC-10. We evaluate different signal functions and reference models.

	Loss				Conf				GN			
Reference Model	Calibrated Acc	Our Acc	Calibrated AUC	Our AUC	Calibrated Acc	Our Acc	Calibrated AUC	Our AUC	Calibrated Acc	Our Acc	Calibrated AUC	Our AUC
Trivial Model [57]	59.6%	**66.6%**	0.658	**0.739**	60.6%	**62.6%**	0.636	**0.687**	61.9%	**66.9%**	0.662	**0.746**
Distilled Model [62]	59.4%	**66.4%**	0.654	**0.749**	56.8%	**62.1%**	0.604	**0.702**	63.3%	**67.1%**	0.680	**0.756**

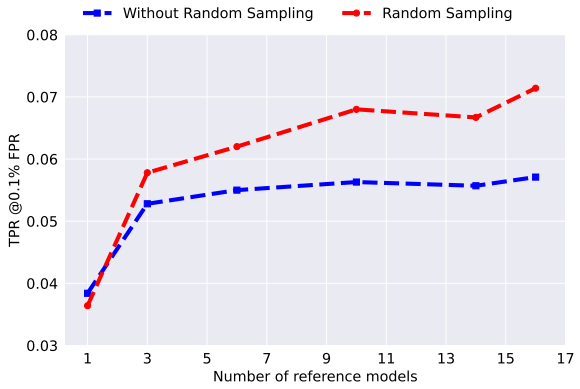


Figure 7: The attack performance exhibits apparent enhancement as the number of reference models increases, with diminishing returns. When using random sampling, the attack results have a higher upper bound.

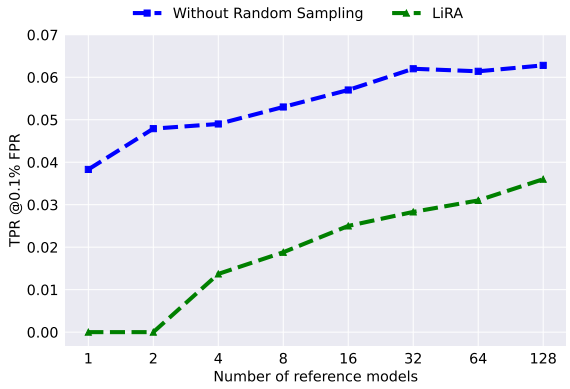


Figure 8: Both RAPID and LiRA exhibit enhanced attack performance as the number of reference models increases, with RAPID benefiting more from training additional reference models.

of 0.1% as the number of $\mathcal{M}_{\text{reference}}$ increases. As expected, a larger number of reference models leads to better attack performance. It is further enhanced when random sampling is employed, as having a larger number of seen data points when training $\mathcal{M}_{\text{reference}}$ means the extent to the target point represented in $\mathcal{D}_{\text{reference}}$, becomes more like that under the entire distribution π. Training more than two reference models brings diminishing benefits as the averaged results gradually stabilize. Another question is whether the rate at which RAPID's attack success rate increases, relative to the associated attack cost, outpaces existing methods, and we use LiRA as the baseline to answer this question. Figure 8 demonstrates that RAPID benefits more from the ability to train increasing numbers of reference models. Specifically, LiRA requires training at least 32 reference models to capture the majority of the benefits.

6.3 Multiple Queries

Previous work [5] has suggested that models are typically trained to minimize their loss on augmented versions of examples, which inspires the idea of conducting MIAs on augmented versions of examples that have been seen during training. However, the results of attacks against CIFAR-10 in Figure 9 are not entirely consistent with this explanation. Note that we average the membership scores obtained from multiple queries on the target model to obtain the final $\mathcal{S}(x, y)$. The experimental results show that increasing the number of queries leads to diminishing improvements in attack performance. Under the previous explanation, the results of each query on different augmented versions of target samples should be independent and of equal importance, so that averaging results of multiple queries should not lead to such significant improvements. The final experimental results actually align perfectly with the analysis provided in Section 4.3. This is also why the trade-off observed with multiple queries is similar to that of random sampling. Furthermore, Figure 9 demonstrates that RAPID continues to outperform LiRA even with an increasing number of queries. This justifies our claim that averaging outputs is equally good for fitting multiple-dimensional spherical Gaussians. Querying the target model only four times can capture the majority of the benefits, which enhances the feasibility of the RAPID attack.

6.4 Model Architecture

Most existing works [5, 29, 31, 41, 46, 62, 63] have assumed that the adversary has knowledge of the specific architecture of the target model in order to have a larger attack surface. However, this assumption is often not valid. Therefore, we aim to investigate the impact of mismatched model architectures on the attack results. Following the settings of our main experiments, we vary the architectures of the target model, shadow models, and reference models while ensuring consistency in the architectures of shadow models and reference models (which is feasible for the attacker). The experimental results in Figure 10 show that the attack performs best when all three models have identical architectures. When the model architectures are completely different, there is only a little

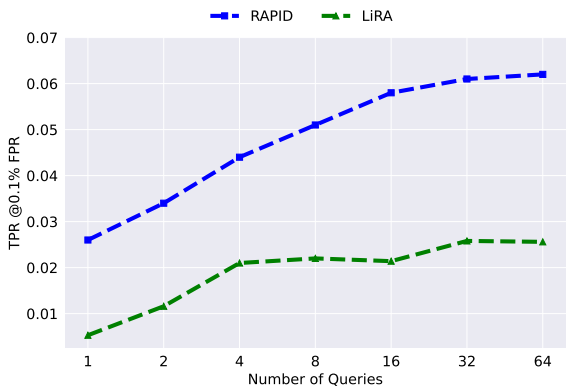


Figure 9: Multiple queries on the augmented versions of the target sample can significantly enhance the attack performances, and RAPID consistently outperforms LiRA.

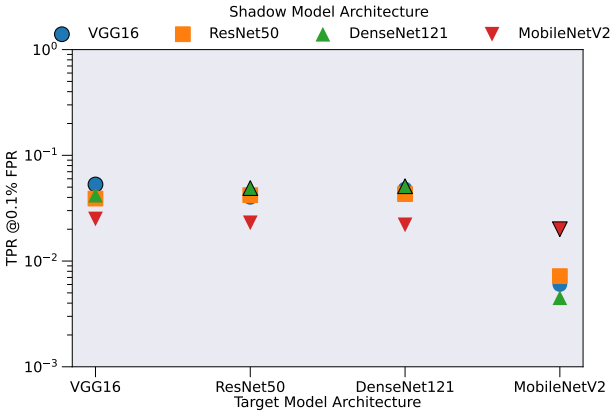


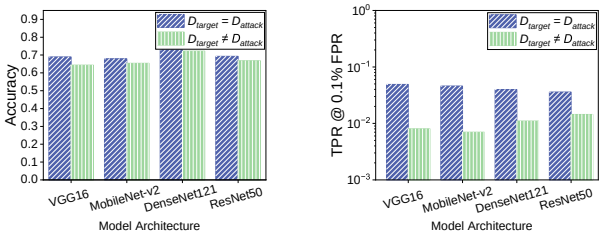
Figure 10: The impact of architecture differences between the target model and the models trained by the adversary (shadow model and reference models) on CINIC-10.

drop in attack performance except for MobileNetV2. The phenomenon of degradation is easily understood because the membership score distributions obtained from models with different architectures are significantly different, even if they are trained on the same dataset. This directly leads to our trained $\mathcal{M}_{\text{score}}$ incorrectly mapping $\mathcal{S}(x, y)$ and $\mathcal{S}'(x, y)$ obtained from $\mathcal{M}_{\text{target}}$ and $\mathcal{M}_{\text{refernce}}$ to final membership scores. Despite that, our attack still achieves significantly better performance compared to other baseline attacks using the same architecture, as demonstrated in Table 2. The notable decrease in attack performance due to the MobileNetV2 architecture can be attributed to the fact that in MobileNetV2, the number of channels in the feature map increases and then decreases, which is contrary to the other three architectures. Remarkably, previous research [5] has also shown similar experimental results. We hope that future work can provide a clearer explanation for this phenomenon. Overall, our attack demonstrates stronger robustness because it outperforms existing baseline attacks even in more challenging settings, whereas the baseline attacks achieved their results in easier settings.

6.5 Disjoint Dataset

In this section, we relax the assumption that the attacker has access to an attack dataset that follows the same distribution as the target model's training dataset. We instead assume that the attacker only has an attack dataset that is disjoint from the target model's training dataset, which they use to train the shadow models and reference models. This is a more realistic condition since it is difficult for the attacker to obtain a dataset that is perfectly aligned with the target training dataset. Specifically, we conduct experiments in the following two settings:

- $\mathbb{D}_{\text{target}} = \mathbb{D}_{\text{attack}}$. Specifically, we train the target model, shadow model, and reference models using the CIFAR-10 dataset. This setup completely follows the settings in our main experiments.
- $\mathbb{D}_{\text{target}} \neq \mathbb{D}_{\text{attack}}$. Specifically, we train the target model using a subset of the CIFAR-10 dataset, while we train the shadow model and reference models using the ImageNet portion of the CINIC-10 dataset, following prior work [5, 31].



(a) Balanced Accuracy **(b) TPR @ 0.1% FPR**

Figure 11: The impact of distribution shift between the target model training dataset and the attack dataset owned by the adversary (shadow dataset and reference dataset).

In order to eliminate the influence of overfitting on the attack performance, we keep the same amount of data in both settings. Additionally, the number of queries and reference models remains the same. Figure 11 shows that the distribution shift between $\mathcal{D}_{\text{target}}$ and $\mathcal{D}_{\text{attack}}$ indeed leads to a noticeable decrease in TPR at 0.1% FPR. This is because the decreasing similarity between the shadow model and the target model makes errors in calibrated scores increase, which finally weakens the performance of scoring model on an unseen dataset. Remarkably, our attack still outperforms the majority of baseline attacks in harder settings on Balanced Accuracy.

7 RELATED WORK

Recently researchers have paid growing emphasis on the importance of high-precision inference in the field of MIAs [5, 29, 31, 33, 41, 57, 58, 62]. Various attack methods based on difficulty calibration have been proposed to address this challenge. Sablayrolles et al. [41] introduce a method that uses loss from both reference models trained with and without the target point to calibrate the original membership scores. Watson et al. [57] employ a similar

Table 8: The attack performances of LiRA online version and our RAPID on VGG16 models trained on four benchmark datasets. We use 64 reference models (only OUT models) for RAPID to achieve its optimal performance.

Attack	TPR @ 0.1% FPR				AUC				Balanced Accuracy			
method	CIFAR-10	CIFAR-100	CINIC-10	SVHN	CIFAR-10	CIFAR-100	CINIC-10	SVHN	CIFAR-10	CIFAR-100	CINIC-10	SVHN
Carlini et al. [5]	**11.9%**	**43.6%**	12.4%	**6.5%**	0.790	0.972	0.778	0.629	68.9%	90.1%	63.9%	57.4%
Ours	10.9%	42.3%	**13.9%**	5.8%	**0.808**	**0.974**	**0.826**	**0.641**	**70.5%**	**90.6%**	**71.8%**	**57.9%**

approach but replace all reference models with trivial OUT models (trained without the target point). Carlini et al. [5] take a step further from the aforementioned approaches [41, 57] by fitting Gaussians to the outputs of the referenced models. It considers the distribution parameters of the target point's loss on a large number of reference models. Ye et al. [62] design a model-dependent and sample-dependent attack leveraging distilled models, which are closer to the target model. Liu et al. [31] introduce a Loss Trajectory Attack, which utilizes the distillation trajectory of the target model for membership inference. Wen et al. [58] argue that one limitation of LiRA is that it queries the target model using only the original target data point or its augmentations. They instead learn query vectors that are maximally discriminative; they separate all models trained with the target data point from all models trained without it. In general, previous work has mainly focused on obtaining calibrated scores of higher-quality, while overlooking the impact of attack cost on the practical threat of the attack. Notably, pioneering work by Shokri et al. [46] also utilizes original outputs, but is primarily based on the intuition that there are differences between outputs from members and non-members. To the best of our knowledge, we are the first to formally utilize the compelling non-member evidence in original outputs to address the inherent errors in difficulty calibration, thereby achieving a more powerful and practical MIA.

8 DISCUSSION AND LIMITATIONS

Our Paradigm Also Enhances LiRA. In addition, our extensive experimental results have demonstrated that RAPID significantly outperforms existing state-of-the-art attacks. However, whether directly re-leveraging original membership scores can enhance more complicated attacks that utilize difficulty calibration (e.g., LiRA) remains an unresolved question. To investigate this question, we do experiments on a VGG16 model trained on the CIFAR-10 dataset. Specifically, we use the concatenation of the original membership scores and the membership scores calculated by LiRA offline (i.e., the results of the one-sided hypothesis test) as features to train a scoring model. The experimental results indicate a significant improvement in LiRA's performance, elevating the TPR @0.1% FPR from 2.2% to 4.5%, the AUC from 0.534 to 0.775, and the Acc from 57.5% to 68.7%. In other words, although LiRA has trained numerous reference models to make a Gaussian likelihood estimate, it still potentially misclassifies certain high-loss non-members as members.

RAPID vs. LiRA Online. As LiRA online is computationally infeasible, we do not include it as a baseline in our main experiments. However, Carlini et al. [5] provide a clever method to evaluate the theoretical performance of LiRA online. We are interested in whether RAPID could serve as a practical alternative to LiRA online in real-world scenarios. Specifically, they combine members and non-members into a set, and then randomly sample half of the data to train a reference model. This process is repeated 256 times. For any given target sample, since it has a 50% chance of being sampled into the training set of any reference model, there are approximately 128 reference models serving as its IN models, and 128 reference models serving as its OUT models. Although this method circumvents the necessity to train 128 IN models for each target sample, it remains impractical in reality because the adversary cannot access all potential member samples before the inference time. Furthermore, this implementation potentially boosts LiRA online's attack performance compared to traditional methods (i.e., those used in our main experiments)—each IN/OUT model shares about half of its training data with the target model, making the IN/OUT models highly similar to the target model. This similarity makes calibrated membership scores more accurate, as they rely more on membership status rather than model parameters (model characteristics). To ensure a fair comparison, we train the reference models for RAPID using a similar method to LiRA. However, note that RAPID remains a practical offline attack, with the key difference being that the reference models and the target model share some training data. As demonstrated in Table 8, our RAPID, as an offline attack, achieves nearly the same TPR at 0.1% FPR as LiRA online, along with higher AUC and balanced accuracy results. Consequently, while LiRA online is not so practical in real-world scenarios, our RAPID represents an equivalent alternative that is computationally feasible.

Limitations. Our work has several limitations. First, the effectiveness of all existing MIAs mainly relies on identifying out-of-distribution member samples (i.e., samples only receiving high membership scores from the target model). This, to some extent, limits the performance of MIAs. Although our RAPID achieves state-of-the-art performance, it does not fully address this issue. Second, we only evaluate RAPID on public datasets, and its effectiveness and sensitivity to specific populations (or subgroups of datasets) have not been fully investigated. Third, our evaluation on LLMs is limited to masked language models, and the performance of RAPID on autoregressive language models has not been studied. We will conduct experiments on more LLMs in the future. Despite these limitations, we believe our study provides insight into the limitations of difficulty calibration, issues of MIA practicality, and finally the potential solutions to the aforementioned issues.

9 CONCLUSION

In this paper, we have emphasized that existing reference-based MIAs do not fully utilize the non-member evidence contained in the

original membership scores, which can be re-leveraged to correct the misclassification of non-members caused by difficulty calibration. Therefore, we have introduced a new attack RAPID, which directly corrects the inherent errors in difficulty calibration by training a scoring model to map the original membership scores and the calibrated scores to the final membership scores. This improves the attack efficacy by eliminating the need for: 1) training a large number of models; and 2) near-unlimited query access to the target model. Extensive experiments demonstrate the state-of-the-art performance of RAPID in both classic image domains and recent fields of LLMs. We hope our research can advance the development of more efficacious techniques for quantifying privacy loss and protecting data privacy.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments. This work is supported in part by the National Natural Science Foundation of China under Grants No. 61872430, 61402342, and 61772384. Additionally, it is sponsored by funding from Research on Attacks and Defenses in Split Learning by Ant Group, China. Any opinions, findings, and conclusions expressed in this paper are those of the authors only and do not necessarily reflect the views of any funding agencies.

REFERENCES

[1] 2006. Hospital discharge data public use data file. (2006).
[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.
[3] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 141–159.
[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[5] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1897–1914.
[6] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting Training Data from Large Language Models.. In *USENIX Security Symposium*, Vol. 6.
[7] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 2020. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 343–362.
[8] Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. 2019. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2287–2295.
[9] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *International conference on machine learning*. PMLR, 1964–1974.
[10] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501* (2018).
[11] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. 2018. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505* (2018).
[12] Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, Vol. 23. 107–124.
[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[14] Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, and Franziska Boenisch. 2023. On the Privacy Risk of In-context Learning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.
[16] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.
[17] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/p18-1082
[18] Jamie Hayes, Luca Melis, George Danezis, and ED Cristofaro. 2017. Logan: Evaluating information leakage of generative models using generative adversarial networks. *arXiv preprint arXiv:1705.07663* 18 (2017).
[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[20] Xinlei He, Zheng Li, Weilin Xu, Cory Cornelius, and Yang Zhang. 2022. Membership-Doctor: Comprehensive Assessment of Membership Inference Against Machine Learning Models. *arXiv preprint arXiv:2208.10445* (2022).
[21] Xinlei He, Rui Wen, Yixin Wu, Michael Backes, Yun Shen, and Yang Zhang. 2021. Node-level membership inference attacks against graph neural networks. *arXiv preprint arXiv:2102.05429* (2021).
[22] Yu He, Boheng Li, Yao Wang, Mengda Yang, Juan Wang, Hongxin Hu, and Xingyu Zhao. 2024. Is Difficulty Calibration All We Need? Towards More Practical Membership Inference Attacks. *arXiv preprint arXiv:2409.00426* (2024).
[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
[24] Ziheng Huang, Boheng Li, Yan Cai, Run Wang, Shangwei Guo, Liming Fang, Jing Chen, and Lina Wang. 2023. What can Discriminator do? Towards Box-free Ownership Verification of Generative Adversarial Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5009–5019.
[25] Kalpesh Krishna, Gaurav Singh Tomar, Ankur P Parikh, Nicolas Papernot, and Mohit Iyyer. 2019. Thieves on sesame street! model extraction of bert-based apis. *arXiv preprint arXiv:1910.12366* (2019).
[26] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
[27] Anders Krogh and John Hertz. 1991. A simple weight decay can improve generalization. *Advances in neural information processing systems* 4 (1991).
[28] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 656–672.
[29] Klas Leino and Matt Fredrikson. 2020. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *29th USENIX Security Symposium*.
[30] Zheng Li and Yang Zhang. 2021. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 880–895.
[31] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. 2022. Membership inference attacks by exploiting loss trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2085–2098.
[32] Yunhui Long, Vincent Bindschaedler, and Carl A Gunter. 2017. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136* (2017).
[33] Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. 2020. A pragmatic approach to membership inferences on machine learning models. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 521–534.
[34] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
[35] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership Inference Attacks against Language Models via Neighbourhood Comparison. *arXiv preprint arXiv:2305.18462* (2023).
[36] H Brendan McMahan, Galen Andrew, Ulfar Erlingsson, Steve Chien, Ilya Mironov, Nicolas Papernot, and Peter Kairouz. 2018. A general approach to adding differential privacy to iterative training procedures. *arXiv preprint arXiv:1812.06210* (2018).
[37] Sasi Kumar Murakonda and Reza Shokri. 2020. ML Privacy Meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. *arXiv preprint arXiv:2007.09339* (2020).
[38] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. 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, 739–753.

[39] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. (2011).

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[41] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*. PMLR, 5558–5567.

[42] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246* (2018).

[43] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[44] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. 2009. Genomic privacy and limits of individual detection in a pool. *Nature genetics* 41, 9 (2009), 965–967.

[45] Shuo Shao, Yiming Li, Hongwei Yao, Yiling He, Zhan Qin, and Kui Ren. 2025. Explanation as a Watermark: Towards Harmless and Multi-bit Model Ownership Verification via Watermarking Feature Attribution. In *NDSS*.

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

[47] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[48] Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 196–206.

[49] Liwei Song and Prateek Mittal. 2021. Systematic Evaluation of Privacy Risks of Machine Learning Models.. In *USENIX Security Symposium*, Vol. 1. 4.

[50] Shuang Song and David Marn. [n. d.]. Introducing a new privacy testing library in tensorflow (2020). *URL https://blog. tensorflow. org/2020/06/introducing-new-privacy-testing-library. html* ([n. d.]).

[51] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2018. Towards demystifying membership inference attacks. *arXiv preprint arXiv:1807.09173* (2018).

[52] David A Van Dyk and Xiao-Li Meng. 2001. The art of data augmentation. *Journal of Computational and Graphical Statistics* 10, 1 (2001), 1–50.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[54] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).

[55] Zhenting Wang, Chen Chen, Yi Zeng, Lingjuan Lyu, and Shiqing Ma. 2024. Where did i come from? origin attribution of ai-generated images. *Advances in neural information processing systems* 36 (2024).

[56] Zhenting Wang, Vikash Sehwag, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. 2024. How to Trace Latent Generative Model Generated Images without Artificial Watermark?. In *Forty-first International Conference on Machine Learning*.

[57] Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. 2021. On the importance of difficulty calibration in membership inference attacks. *arXiv preprint arXiv:2111.08440* (2021).

[58] Yuxin Wen, Arpit Bansal, Hamid Kazemi, Eitan Borgnia, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2022. Canary in a Coalmine: Better Membership Inference with Ensembled Adversarial Queries. In *The Eleventh International Conference on Learning Representations*.

[59] Yutong Wu, Han Qiu, Shangwei Guo, Jiwei Li, and Tianwei Zhang. 2024. You Only Query Once: An Efficient Label-Only Membership Inference Attack. In *The Twelfth International Conference on Learning Representations*.

[60] Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. 2015. Nationtelescope: Monitoring and visualizing large-scale collective behavior in lbsns. *Journal of Network and Computer Applications* 55 (2015), 170–180.

[61] Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 3 (2016), 1–23.

[62] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 3093–3106.

[63] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 268–282.

[64] Xiaoyong Yuan and Lan Zhang. 2022. Membership inference attacks and defenses in neural network pruning. In *31st USENIX Security Symposium (USENIX Security 22)*. 4561–4578.

[65] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. 2021. Membership inference attacks against recommender systems. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 864–879.

[66] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13001–13008.