Empirical Comparison of Membership Inference Attacks in Deep Transfer Learning

Yuxuan Bai [1] **Gauri Pradhan** [1] **Marlon Tobaben** [1] **Antti Honkela** [1]

Abstract

With the emergence of powerful large-scale foundation models, the training paradigm is increasingly shifting from from-scratch training to transfer learning. This enables high utility training with small, domain-specific datasets typical in sensitive applications. Membership inference attacks (MIAs) provide an empirical estimate of the privacy leakage by machine learning models. Yet, prior assessments of MIAs against models fine-tuned with transfer learning rely on a small subset of possible attacks. We address this by comparing performance of diverse MIAs in transfer learning settings to help practitioners identify the most efficient attacks for privacy risk evaluation. We find that attack efficacy decreases with the increase in training data for score-based MIAs. We find that there is no one MIA which captures all privacy risks in models trained with transfer learning. While the Likelihood Ratio Attack (LiRA) demonstrates superior performance across most experimental scenarios, the Inverse Hessian Attack (IHA) proves to be more effective against models fine-tuned on PatchCamelyon dataset in high data regime.

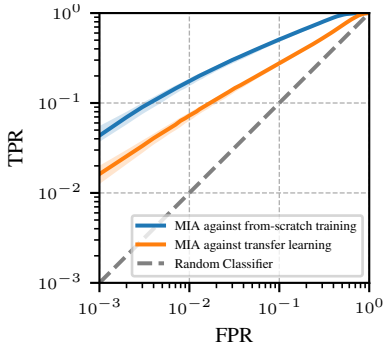

Figure 1. MIA efficacy as measured using LiRA (Carlini et al., 2022) against WideResNet-50-2 (Zagoruyko & Komodakis, 2016) trained from scratch using CIFAR-10 versus the same model pre-trained on ImageNet-1k (Deng et al., 2009) when only the last linear layer of the model is fine-tuned on CIFAR-10. The results are averaged over 3 repeats, with each repeat using $M + 1$ target models ($M = 64$) that share the same optimized hyperparameters obtained through hyperparameter optimization (HPO). The errorbars represent the interquartile range (IQR) of corresponding TPR at FPR. The plot demonstrates that the attack does not behave similarly across the 2 training paradigms, highlighting the need to investigate the performance of different MIA approaches against foundation models fine-tuned using deep transfer learning to ensure that a strong attack is used to evaluate their privacy risks.

1. Introduction

As foundation models increasingly power modern AI systems, their adaptation through transfer learning raises privacy concerns. Recent research has demonstrated that fine-tuned models can inadvertently memorize their training data rather than learning generalizable patterns (Chu et al., 2025), creating potential privacy vulnerabilities.

Membership inference attacks (MIAs) (Shokri et al., 2017) have emerged as a critical tool for quantifying such privacy leakage by determining whether specific data points were used during model training. These attacks not only provide empirical lower bounds on privacy guarantees of a training algorithm, but also expose privacy vulnerabilities in model training strategies. Despite significant advances in MIA methodologies, their evaluation has predominantly focused on models trained from scratch. Figure 1 shows the varying privacy vulnerabilities exploitable by MIAs between fine-tuned and from-scratch trained models, even when both are trained on identical datasets. This divergence suggests that MIA efficacy in transfer learning fundamentally differs from that in from-scratch training.

Earlier works such as Carlini et al. (2019; 2021); Lee et al. (2022); Kandpal et al. (2022) mainly focused on examining memorization of pre-training data by large foundation models. Meeus et al. (2025) recommend using fine-tuned versions of large language models to evaluate memoriza-

[1]Department of Computer Science, University of Helsinki, Finland. Correspondence to: Antti Honkela <antti.honkela@helsinki.fi>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models.

tion using MIAs but do not conduct any experiments in this setting. Other studies using MIAs in transfer learning (Tobaben et al., 2023; 2024; Pradhan et al., 2025) have typically used a limited selection of attacks. A related study on the efficacy of MIAs against machine unlearning by Hayes et al. (2025) shows that the use of weaker versions of MIA overestimates the privacy protection provided by existing unlearning techniques. Thus, it is crucial to establish the relative efficacy of different MIAs to ensure that weaker attacks, which underestimate privacy risks, are not used to evaluate membership privacy in transfer learning scenarios.

Our Contributions In this work, we conducted a systematic evaluation of existing score-based MIAs (Yeom et al., 2018; Salem et al., 2019; Ye et al., 2022; Liu et al., 2022; Bertran et al., 2023; Li et al., 2024; Suri et al., 2024) in transfer learning contexts. Motivated by the work by Tobaben et al. (2024), we investigate the relationship between MIA efficacy and fine-tuning dataset size using consistent experimental setups. Our results confirm that MIA efficacy generally decreases as the number of examples per class increases for most score-based attacks in transfer learning, consistent with the power-law relationship previously observed. However, we identify a notable exception: the white-box Inverse Hessian Attack (IHA) (Suri et al., 2024) exhibits markedly different behavior, demonstrating superior performance in high-shot regimes on PatchCamelyon compared to black-box methods. Additionally, we analyze the effects of changing the training paradigm and properties of the attacks on MIA efficacy.

2. Related Work

Deep Transfer Learning Deep transfer learning has been widely adopted in machine learning, leveraging knowledge from source tasks to enhance performance on target tasks with limited data (Yosinski et al., 2014). The process involves pre-training on large-scale datasets to learn general-purpose feature representations, followed by fine-tuning on smaller, task-specific datasets, reducing data requirements and computational costs. However, this approach introduces privacy vulnerabilities. Models may memorize patterns from source datasets (Tramèr et al., 2024). Additionally, fine-tuning on small target datasets often leads to overfitting, increasing vulnerability to privacy attacks that can extract information about individual training samples. Researchers commonly use pre-trained models like ResNet (He et al., 2016; Kolesnikov et al., 2020) and Vision Transformer (ViT) (Dosovitskiy et al., 2021) due to computational constraints. Therefore, evaluating and mitigating privacy leakage during fine-tuning on sensitive downstream tasks forms a practical motivation for privacy research.

Membership Inference Attacks Membership inference attacks (MIAs) aim to determine whether a specific data sample was used in training dataset of a target model. These attacks exploit differences in model behavior when responding to samples used for training the model (member samples) versus non-member samples, thereby compromising the privacy of sensitive data. MIAs are typically categorized based on the adversary's knowledge and access to the target model (Hu et al., 2022b). In the white-box setting, attackers have full access to the model's learned parameters, gradients, and architecture details. In contrast, black-box attacks operate with limited information, typically requiring knowledge of the data distribution and potentially access to model architecture and hyperparameters. Black-box MIAs further diverge into 2 primary variants: *score-based MIAs* that exploit the model's confidence scores, and *label-only MIAs* that function only with the predicted class labels (Li & Zhang, 2021; Choquette-Choo et al., 2021; Peng et al., 2024). Our work mainly focuses on score-based MIAs, as they represent the most potent yet practically feasible attacks.

Shadow-model-based vs. Shadow-model-free MIAs Score-based MIAs can be further divided into shadow-model-based and shadow-model-free approaches. Shadow-model-based MIAs depend on *shadow training* (Shokri et al., 2017), a technique where the attacker trains surrogate models that mimic the behaviour of the target model. These include methods such as ML-Leaks (Adversary 1) (Salem et al., 2019), Trajectory-MIA (Liu et al., 2022), Sequential-Metric based MIA (SeqMIA) (Li et al., 2024), Likelihood Ratio Attack (LiRA) (Carlini et al., 2022), and Robust MIA (RMIA) (Zarifzadeh et al., 2024). Despite their effectiveness, shadow-model-based MIAs require substantial computational resources, especially since their attack efficiency relies on training additional models. This limitation has motivated the development of more computationally efficient shadow-model-free alternatives, including LOSS attack (Yeom et al., 2018), Attack-P (Ye et al., 2022), and quantile-MIA (QMIA) (Bertran et al., 2023).

Black-box vs. White-box MIAs While most MIA research has focused on black-box settings following the theoretical assertion by Sablayrolles et al. (2019) stating that white-box access provides no additional advantage for evaluating membership privacy, Suri et al. (2024) recently challenged this consensus by deriving a new optimality condition and introducing the white-box Inverse Hessian Attack (IHA). For completeness, our evaluation includes this white-box approach alongside predominant black-box methods.

Table 1 summarizes the threat models for all MIAs used in this paper.

Table 1. Summarizing MIAs in terms of the auxiliary information about the target model available to the attacker.

Attack	Target Model Access			
	Data Distribution	Architecture	Hyperparameters	Model Parameters
LOSS (Yeom et al., 2018)	✓	-	-	-
Attack-P (Ye et al., 2022)	✓	-	-	-
QMIA (Bertran et al., 2023)	✓	-	-	-
LiRA (Carlini et al., 2022)	✓	✓	✓	-
RMIA (Zarifzadeh et al., 2024)	✓	✓	✓	-
ML-Leaks (Salem et al., 2019)	✓	✓	✓	-
Trajectory-MIA (Liu et al., 2022)	✓	✓	✓	-
IHA (Suri et al., 2024)	✓	✓	✓	✓

3. Score-based MIAs

In this section, we describe the different MIAs employed in this paper and highlight how they differ in their approach to estimate membership privacy.

Preliminaries Let $\mathcal{D}$ be a dataset sampled from data distribution π. This dataset is used to train a machine learning model $\mathcal{M}$ with parameters θ. Next, we establish the probability notations used in the paper. $\Pr(\theta|x)$ denotes the probability of observing parameters θ when x in included in the training set, while $\Pr(\theta|\overline{x})$ denotes the probability of observing θ if x is *not* in the training set. Conversely, $\Pr(x|\theta)$ represents the probability that x was part of the training set that produced θ.

3.1. Shadow-model-based MIAs

ML-Leaks ML-Leaks (Adversary 1) (Salem et al., 2019) refines the shadow training approach (Shokri et al., 2017). It begins by training a shadow model using the dataset $\mathcal{D}_{\text{shadow}}$ sampled from π. However, instead of using the full prediction vector, ML-Leaks extracts only the top 3 posterior probabilities (or top 2 for binary-class datasets), ordered from highest to lowest for each sample in $\mathcal{D}_{\text{shadow}}$. Following this, it uses the trimmed probability vectors as inputs to train the attack model. This attack model can then be deployed to compute the membership scores for samples in $\mathcal{D}$.

LiRA LiRA is a hypothesis testing framework for membership inference proposed by Carlini et al. (2022). For a given target model, it trains M shadow models, such that the target sample, x, is included in the training dataset for $1/2$ of them (IN models) whereas it is excluded from the training dataset of the remaining $M/2$ models (OUT models). Using the predicted and logit-scaled confidence scores from these IN and OUT shadow models, the attacker can build the IN and OUT Gaussian distributions. Following this, the attacker can employ a likelihood ratio (LR) test (Neyman & Pearson, 1933) to compare $\Pr(\theta|x)$ against $\Pr(\theta|\overline{x})$:

$$\text{LR}_\theta(x) = \frac{\Pr(\theta|x)}{\Pr(\theta|\overline{x})}. \tag{1}$$

The attacker can use $\text{LR}_\theta(x)$ as the membership score to train a binary classifier to differentiate between member and non-member samples.

Trajectory-MIA Liu et al. (2022) proposed Trajectory-MIA as an efficient alternative to LiRA. This is because Trajectory-MIA uses a single shadow model compared to LiRA's multiple shadow models' based approach. The method leverages knowledge distillation (Hinton et al., 2015) to simulate the target model's training trajectory. Specifically, it performs distillation on both the target and shadow models, minimizing Kullback-Leibler (KL) divergence between the student and teacher models' outputs. By recording the per-example training loss trajectory across distillation epochs, Trajectory-MIA captures temporal patterns that differ between member and non-member samples. These loss trajectories serve as feature vectors for a MLP classifier to predict memberships.

RMIA RMIA (Zarifzadeh et al., 2024) further refines LiRA by incorporating knowledge about the population data $z \sim \pi$ in the likelihood ratio. It introduces a pairwise LR test that explicitly incorporates population samples z, computing the probability that the pairwise LR exceeds a preset threshold γ:

$$\Pr_{z\sim\pi}(\text{LR}_\theta(x,z) \geq \gamma) = \Pr_{z\sim\pi}(\frac{\Pr(\theta|x)}{\Pr(\theta|z)} \geq \gamma). \tag{2}$$

This pairwise LR formulation captures the relative relationship between a potential member sample x and known non-member samples z drawn from the population. By composing these pairwise comparisons, Zarifzadeh et al. (2024) contend that RMIA achieves greater robustness to distribution shifts between members and non-members.

3.2. Shadow-model-free MIAs

LOSS Attack/ Attack-P LOSS attack (Yeom et al., 2018) uses loss on the target sample $\ell(\mathcal{M}(x), y)$ as a membership signal. Since the objective of training a machine learning model is usually to minimize their loss on the training samples, it compares the loss on the target sample against a fixed threshold τ to infer membership. Attack-P (Ye et al., 2022) is an improvised version of the LOSS attack, which constructs an empirical cumulative distribution function (CDF) from the population samples. Unlike the LOSS attack which uses a fixed threshold τ, Attack-P compares the loss of the target sample to the distribution of losses from known non-members, calculating what percentage of non-member losses fall below the target's loss. Specifically, these attacks compare $\Pr(x|\theta)$ with $\Pr(z|\theta)$ as the threshold to determine the membership of x.

QMIA QMIA (Bertran et al., 2023) determines per-sample thresholds by performing quantile regression on the distribution of confidence scores obtained from known non-member data. By training a regression model with pinball loss to predict these thresholds at a desired false positive rate, QMIA creates a nuanced decision boundary that adapts to individual sample. This model-agnostic approach was proposed as an effective alternative to shadow-model-based MIAs in black-box settings where only API access is available.

IHA Suri et al. (2024) proposed a white-box variant of MIA, namely the Inverse Hessian Attack (IHA). Building on recent advancements in discrete-time SGD dynamics (Liu et al., 2021; Ziyin et al., 2022), Suri et al. (2024) demonstrate that an optimal membership inference requires white-box access to the model's parameters post-training, and not merely its output predictions. IHA relies on a local similarity assumption, which posits that models trained with or without a specific data point converge to similar local minima. Under this assumption, the Hessian matrices at the respective optima share similar structure: $\mathbf{H}_* = \mathbf{H}_0(\boldsymbol{w}_0^*) = \mathbf{H}_1(\boldsymbol{w}_1^*)$, where $\boldsymbol{w}_0^*$ and $\boldsymbol{w}_1^*$ represent the optimal parameters for models trained without and with the target sample, respectively. In addition, the loss functions achieve similar values at these local minima: $L_* = L_0(\boldsymbol{w}_0^*) = L_1(\boldsymbol{w}_1^*)$. This assumption allows IHA to approximate the optimal membership inference by formulating the MIA scoring function using terms dependent on gradients and model parameters.

4. Experimental Setup

Datasets We use CIFAR-10, CIFAR-100 (Krizhevsky, 2009), and PatchCamelyon (Veeling et al., 2018) in our experiments. CIFAR-10 and CIFAR-100 are common benchmark datasets for MIA evaluation. PatchCamelyon, including only 2 classes, enables experiments with substantially larger number of shots S (examples per class), providing greater insight into how training set size affects MIA efficacy.

Models We use ViT-B/16 (Dosovitskiy et al., 2021) and BiT-M-R50x1 (R-50) (Kolesnikov et al., 2020) as the backbone models for fine-tuning, both pretrained on ImageNet-21k (Deng et al., 2009).

Parameterization We employ 2 schemes for parameterization: *(i) Head-only*, where only the classification layer is replaced by a trainable linear layer, with initial weights set to 0, while the feature extraction backbone remains frozen, and *(ii) FiLM*, where FiLM adapters (Perez et al., 2018) are introduced throughout the network alongside a trainable classification head. This parameter-efficient technique, applicable to both convolutional and transformer architectures, enables more expressive adaptation while minimizing trainable parameters compared to full fine-tuning. Although alternatives such as LoRA (Hu et al., 2022a), and CaSE (Patacchiola et al., 2022) exist, FiLM is selected due to its demonstrated effectiveness in parameter-efficient few-shot transfer learning (Shysheya et al., 2023; Tobaben et al., 2023).

Hyperparameter Optimization Before model training, we perform hyperparameter optimization (HPO) to identify the optimal set of hyperparameters to train the target model. We use the same set of hyperparameters to train both the target and the shadow model(s). We begin by sampling $1/2$ of the training dataset ($\mathcal{D}$) for HPO. In each HPO trial, we use 70% of the data for training the model while the remaining 30% is used as validation dataset. We run the HPO for 20 trials to explore the hyperparameter space. We implement HPO using Optuna (Akiba et al., 2019) with Tree-structured Parzen Estimator (TPE) algorithm (Bergstra et al., 2011). Table 2 summarizes the hyperparameters and their corresponding search ranges used in our experiments.

Table 2. Hyperparameter search ranges used for Bayesian optimization with Optuna.

Hyperparameter	Parameterization	Range
Epoch	Head-only	[1, 200]
	FiLM	[1, 40]
Train Batch Size	Head-only/FiLM	[10, 1000]
Learning Rate	Head-only/FiLM	$[10^{-7}, 10^{-2}]$

Metrics The metrics used for evaluating efficacy of MIAs include: *(i) log-scaled ROC* curves to visualize the trade-off

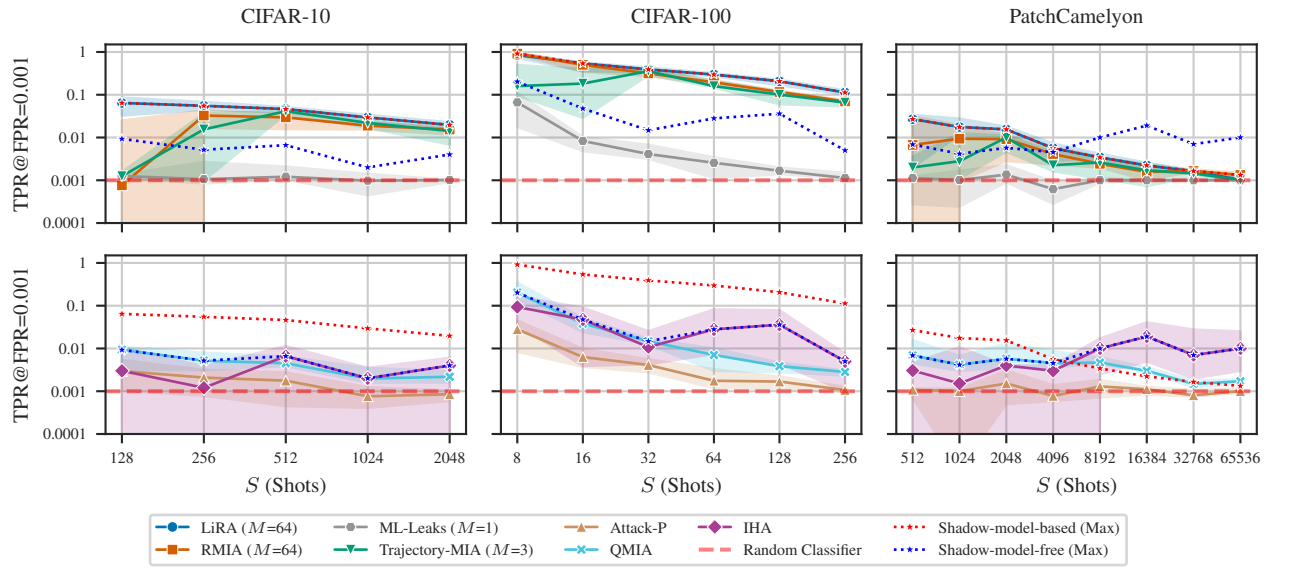

Figure 2. **MIA efficacy against ViT-B/16 (Head-only) models as a function of S (shots).** Upper: Shadow-model-based attacks using M shadow models. Lower: Shadow-model-free attacks. The errorbars represent the interquartile range (IQR) of the estimated TPR at fixed FPR and the dotted lines represent the maximum of the median MIA efficacy of shadow-model-based and shadow-model-free attacks. Shadow-model-based attacks generally demonstrate more stable and stronger MIA efficacy compared to shadow-model-free attacks. In the high-shot regime of PatchCamelyon, however, the white-box IHA has a considerable advantage over other MIAs in terms of MIA efficacy. Results are averaged over 10 repeats and we use 1 target model per repeat.

between true positive rate (TPR) and false positive rate (FPR), with emphasis on the low FPR region, (*ii) TPR at low FPR*: to measure MIA efficacy when the false positives must be minimized, better representing realistic attack scenarios (Carlini et al., 2022), and (*iii) Interquartile Range (IQR)* is used as a statistical measure to quantify uncertainty. It is the difference between the 25^{th} and 75^{th} percentile of a set of values and provides a trimmed estimation that is less sensitive to outliers compared to standard deviation. This helps mitigate the impact of extreme values that may occur due to particularly favorable or unfavorable random initializations, providing a more reliable assessment of variation in MIA efficacy across different experimental repeats.

Experimental Protocol Unless otherwise specified, all our results presented are averaged over 10 experimental repeats. Within each repeat, we sample a new subset of the population dataset (e.g. CIFAR-10). We sample the datasets to train the target and shadow models from the selected subset such that for each sample x in the target model's training dataset, we have $1/2$ of the shadow models trained with x whereas the remaining are not training with x. Additionally, we run the HPO algorithm for each experimental repeat to find optimal hyperparameters to train the target and shadow models. In order to ensure fair comparisons between different MIAs, we restrict our experiments to 1 target model per repeat as more target models are computationally expensive. With LiRA it is possible to use the so-called efficient LiRA implementation proposed by Carlini et al. (2022) that uses every shadow model also as a target model (see Section 5.3). While it is computationally cheap to implement the same also for RMIA, it is computationally expensive to do the same for the other attacks.

5. Results

We comprehensively analyze the factors that influence MIA efficacy of various attacks in transfer learning. In Section 5.1, we explore the effect of the number of shots on the performance of different attacks. In Section 5.2, we study how the performance of attacks varies for different parameterization schemes. In Section 5.3 and Section 5.4, we study the impact of number of shadow models and data augmentation on the most powerful MIAs, namely, LiRA and RMIA.

In addition, we investigate the impact of attack-specific parameters on MIA efficacy by evaluating how choice of attack threshold, γ, for RMIA (Appendix A.1) and distillation set size for Trajectory-MIA (Appendix A.2) affect their respective performance. These attacks have been proposed as more efficient alternatives to LiRA but we find their performance to be sensitive to the choice of hyperparameters, such as γ in RMIA, introduced by the attack design.

5.1. Effect of Training Dataset Properties

Figure 2 demonstrates the relationships between the MIA efficacy and the number of examples per class, or shots (S), across all datasets, confirming the power law relationship proposed by Tobaben et al. (2024). Among all examined attacks, LiRA consistently outperforms other approaches across most experimental settings, exhibiting remarkable stability as evidenced by the narrower confidence intervals for it. The performance advantage of LiRA is particularly pronounced at lower S, though this advantage diminishes as S increases.

While most attacks show monotonic degradation in MIA efficacy with increase in S, IHA displays non-monotonic patterns with multiple fluctuations. This anomalous deviation from the general trend suggests that white-box attacks can exploit different aspects of model vulnerability. The second-order information captured by the Hessian matrix appears to reveal membership signals that might go undetected by the black-box approaches in large training datasets. This demonstrates that while simply increasing training data volume can effectively reduce average vulnerability to existing black-box MIAs, it cannot completely eliminate privacy risks to white-box attacks.

5.2. Effect of Training Paradigm

Figure 3 illustrates the difference in MIA efficacy between FiLM and Head-only fine-tuned R-50 models across multiple MIAs. Overall, parameterization strategies have minimal impact on MIA efficacy across most MIAs. The only notable exception is Trajectory-MIA which shows decreased MIA efficacy for FiLM as compared to Head-only. However, Trajectory-MIA remains a weaker attack than LiRA against FiLM. Given that LiRA maintains consistent performance across both parameterization schemes, this suggests that practitioners should choose parameterization scheme based primarily on utility considerations rather than privacy concerns, as the fine-tuning choice does not significantly alter the efficacy of the most effective MIA approaches.

5.3. Number of Shadow Models

Figure 4 illustrates the relationship between the number of shadow models M and MIA efficacy for LiRA and RMIA, 2 of the strongest shadow-model-based attacks discussed in this paper. To ensure a statistically robust evaluation, we employ an efficient implementation of LiRA and RMIA proposed by Carlini et al. (2022). This approach involves sampling $M+1$ datasets from the training dataset $\mathcal{D}$, which contains $C \times S$ samples (C classes with S examples per class), such that each sample has a 0.5 probability of being selected for any given dataset. We then train models on each of these datasets and evaluate attacks against each model

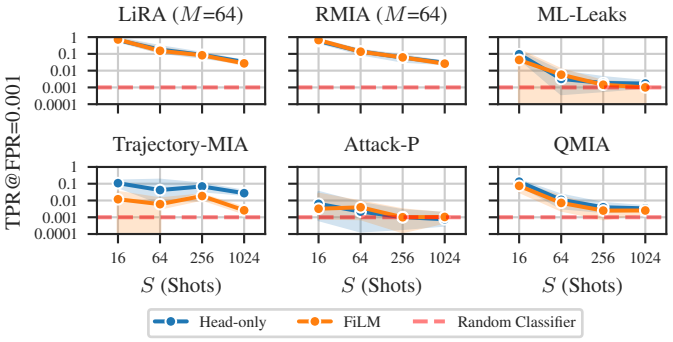

Figure 3. Comparison of MIA efficacy against R-50 fine-tuned on CIFAR-10 with Head-only versus FiLM parameterization across 4 different S (shots). The errorbars represent the interquartile range (IQR) of the estimated TPR at fixed FPR. Results are averaged over 5 repeats with 1 target model in each repeat. For the strongest attacks, there is no considerable difference in MIA efficacy across the 2 parameterization schemes.

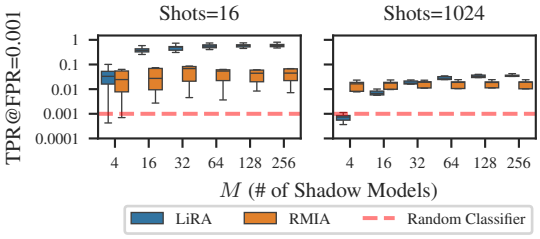

Figure 4. Relationship between MIA efficacy and the number of shadow models (M) for LiRA and RMIA against ViT-B/16 model with Head-only fine-tuned on CIFAR-10. Results demonstrates MIA efficacy in low data availability (shots $S = 16$) and high data availability ($S = 1024$) scenarios. For each configuration, we train $M + 1$ models per repeat, using each model as the target while the remaining M serve as shadow models. We compute the average MIA efficacy (TPR at fixed FPR) across all $M + 1$ target models per repeat, then construct boxplots using these average TPR from 5 independent repeats. LiRA dominates in terms of efficacy over RMIA despite the latter's performance being more robust to the variations in M.

while using the remaining M models as shadow models.

LiRA exhibits significant sensitivity to variations in M across both low-data ($S = 16$) and data-rich ($S = 1024$) scenarios but is robust to increase in M beyond $M \geq 64$. While RMIA demonstrates robustness to changes in M, its efficacy does not exceed LiRA in either of the scenarios. The performance for both the attacks stabilize beyond $M \geq 64$ suggesting it to be a cost-efficient choice for the number of shadow models in deep transfer learning setting.

5.4. Effect of Using Data Augmentation During Fine-Tuning

Prior research has demonstrated that both LiRA and RMIA benefit from querying each sample multiple times when attacking models trained from scratch using data augmentation (Carlini et al., 2022; Zarifzadeh et al., 2024). These attacks achieve improved performance by using not only the original sample but also augmented versions of it during the inference process. To determine whether similar improvements occur in deep transfer learning scenarios, we fine-tune target models using training datasets augmented with simple transformations, including mirror flipping and pixel shifting. This approach is employed in both from-scratch training (Perez & Wang, 2017) and transfer learning (Mehta et al., 2023) to improve model generalization, particularly when working with limited data.

Following the same efficient implementation as described in Section 5.3, we evaluated MIA efficacy using multiple augmented queries generated using a subset of transformations applied during training. For LiRA, the membership signal is averaged over multiple queries directly, while for RMIA, we follow the majority voting scheme as recommended by Zarifzadeh et al. (2024). Figure 5 shows that data augmentation produces negligible performance improvements for both attacks when targeting Head-only fine-tuned models. This finding represents a significant departure from the from-scratch training findings, suggesting different vulnerability patterns in transfer learning. Based on these results, we employ the non-augmented version of both attacks in all experiments for computational efficiency.

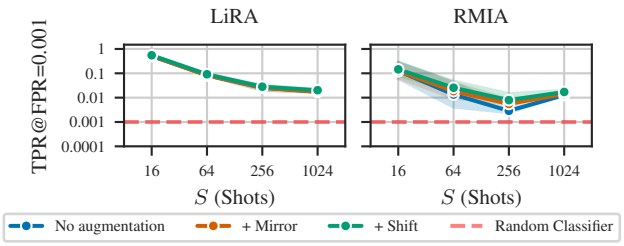

Figure 5. Impact of data augmentation on MIA efficacy across S (shots) for ViT-B/16 models Head-only fine-tuned on CIFAR-10 *with data augmentations*. We compare 2 augmentation strategies: + *Mirror* (where original image plus a horizontally flipped copy of it are used to train the target model) and + *Shift* (where horizontally flipping and/or ± 1-pixel shifts are applied to the original image), with *No augmentation* as the baseline. The errorbars represent the interquartile range (IQR) of the estimated TPR at fixed FPR. Results are averaged over 5 repeats and we use $M + 1$ target models ($M = 64$) per repeat.

6. Discussion

Our findings largely corroborate the claim made by Tobaben et al. (2024) which states that increasing the number of examples per class generally reduces MIA efficacy. However, we do not find this behavior to be consistent across all the attacks. For example, MIA efficacy of IHA increases substantially at larger shots as observed in Figure 2, indicating that white-box attacks can detect vulnerabilities that remain hidden from black-box methods.

For the most powerful attacks (e.g. LiRA) we find that different parameterization schemes, such as Head-only and FiLM, show minimal differences in terms of MIA efficacy. One notable exceptions is Trajectory-MIA, which shows increased vulnerability against Head-only fine-tuning. However, Trajectory-MIA is weaker and less stable compared to LiRA, which maintains consistent performance across both parameterization schemes. This implies that the parameterization choices for fine-tuning could be guided primarily by the utility as they do not significantly affect the performance of the strongest attacks.

Carlini et al. (2022); Zarifzadeh et al. (2024) suggest that data augmentation can be utilized to improve MIA efficacy against models trained from scratch. However, in Section 5.4 we observe no significant improvement in MIA efficacy due to augmentation against models with the last linear layer subject to fine-tuning. This suggests that practitioners can leverage augmentation techniques to improve model utility without substantially compromising its privacy.

No single MIA is able to capture all vulnerabilities in fine-tuned models. LiRA provides robust auditing capabilities but shows decreased efficacy as dataset sizes grow. IHA shows potential to detect vulnerabilities missed by black-box attacks, particularly with datasets that have different characteristics from the pre-training data, such as Patch-Camelyon (Goyal et al., 2023; Choi et al., 2024; Thaker et al., 2024). Comprehensive privacy auditing requires a multi-faceted approach that combines both black-box and white-box methods.

Limitations

- We focus on balanced datasets in our experiments to evaluate the relationship between examples per class (S) and MIA efficacy for different attacks because this design choice enables clear comparison across different data availability scenarios. Future work could extend this analysis to imbalanced datasets commonly found in real-world deployments.

- To ensure a fair comparison across different attacks in Sections 5.1 and 5.2, we restrict our experiments

to having 1 target model per repeat. This differs from Carlini et al. (2022)'s efficient implementation of LiRA where they reuse all the $M + 1$ trained models as the target model and average the MIA efficacy over all of them. This is because attacks such as Trajectory-MIA require training 2 additional distilled models for each target model- shadow model pair, which makes the attack computationally infeasible if the number of target models per repeat is set to be the same as LiRA ($M + 1$). Similar computational constraints are associated with IHA where approximating iHVPs per target sample per model will be prohibitively expensive for large numbers of target models.

7. Conclusion

In this work, we evaluated and compared the performance over a large set of MIAs in transfer learning settings. We found that the attack strength deteriorates as the dataset size increases for black-box MIAs. This agrees with the power law postulated by Tobaben et al. (2024). However, this relationship is not guaranteed to hold for all the attacks discussed in this paper, such as the white-box IHA. This shows that there is no single existing attack that can fully quantify the privacy leakage in deep transfer learning. In addition, we found no significant difference in MIA efficacy between Head-only and FiLM parameterization strategies for most attacks, implying that choice of parameterization can be made with a utility-first perspective. However, MIAs are sensitive to the choice of attack properties, such as the number of shadow models used for the attack. These empirical findings provide guidance for practitioners seeking to assess privacy risks using MIAs in deep transfer learning applications.

Acknowledgments

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI, Grant 356499 and Grant 359111), the Strategic Research Council at the Research Council of Finland (Grant 358247) as well as the European Union (Project 101070617). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. The authors wish to thank the CSC – IT Center for Science, Finland for supporting this project with computational and data storage resources.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631. Association for Computing Machinery, 2019.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for hyper-parameter optimization. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, volume 24, pp. 2546–2554. Curran Associates Inc., 2011.

Bertran, M., Tang, S., Roth, A., Kearns, M., Morgenstern, J. H., and Wu, S. Z. Scalable membership inference attacks via quantile regression. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, volume 36, pp. 314–330, 2023.

Carlini, N., Liu, C., Erlingsson, U., Kos, J., and Song, D. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, pp. 267–284. USENIX Association, 2019.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE Computer Society, 2022.

Choi, C., Lee, Y., Chen, A. S., Zhou, A., Raghunathan, A., and Finn, C. AutoFT: Robust fine-tuning by optimizing hyperparameters on OOD data. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2024.

Choquette-Choo, C. A., Tramer, F., Carlini, N., and Papernot, N. Label-only membership inference attacks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 1964–1974. PMLR, 2021.

Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Levine, S., and Ma, Y. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *The Second Conference on Parsimony and Learning (Recent Spotlight Track)*, 2025.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE Computer Society, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Goyal, S., Kumar, A., Garg, S., Kolter, Z., and Raghunathan, A. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19338–19347, 2023.

Hayes, J., Shumailov, I., Triantafillou, E., Khalifa, A., and Papernot, N. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 497–519. IEEE Computer Society, 2025.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE Computer Society, 2016.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv Preprint arXiv: 1503.02531*, 2015.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Weizhu, C. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022a.

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang, X. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s), 2022b.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 10697–10707. PMLR, 2022.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big Transfer (BiT): General visual representation learning. In *Computer Vision – ECCV 2020*, pp. 491–507. Springer-Verlag, 2020.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445. Association for Computational Linguistics, 2022.

Li, H., Li, Z., Wu, S., Hu, C., Ye, Y., Zhang, M., Feng, D., and Zhang, Y. SeqMIA: Sequential-metric based membership inference attack. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pp. 3496–3510. Association for Computing Machinery, 2024.

Li, Z. and Zhang, Y. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pp. 880–895. Association for Computing Machinery, 2021.

Liu, K., Ziyin, L., and Ueda, M. Noise and fluctuation of finite learning rate stochastic gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7045–7056. PMLR, 2021.

Liu, Y., Zhao, Z., Backes, M., and Zhang, Y. Membership inference attacks by exploiting loss trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, pp. 2085–2098. Association for Computing Machinery, 2022.

Meeus, M., Shilov, I., Jain, S., Faysse, M., Rei, M., and de Montjoye, Y.-A. SoK: membership inference attacks on LLMs are rushing nowhere (and how to fix it). In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 385–401. IEEE Computer Society, 2025.

Mehta, H., Thakurta, A. G., Kurakin, A., and Cutkosky, A. Towards large scale transfer learning for differentially private image classification. *Transactions on Machine Learning Research*, 2023.

Neyman, J. and Pearson, E. S. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical transactions of the Royal Society of London*, 231 (694-706):289–337, 1933.

Patacchiola, M., Bronskill, J., Shysheya, A., Hofmann, K., Nowozin, S., and Turner, R. Contextual squeeze-and-excitation for efficient few-shot image classification. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, volume 35, pp. 36680–36692. Curran Associates, Inc., 2022.

Peng, Y., Roh, J., Maji, S., and Houmansadr, A. Oslo: One-shot label-only membership inference attacks. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pp. 62310–62333. Curran Associates, Inc., 2024.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. FiLM: Visual reasoning with a general

conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Perez, L. and Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Pradhan, G., Jälkö, J., Tobaben, M., and Honkela, A. Hyperparameters in score-based membership inference attacks. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 362–384. IEEE Computer Society, 2025.

Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jégou, H. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5558–5567. PMLR, 2019.

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., and Backes, M. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Proceedings of 26th Annual Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE Computer Society, 2017.

Shysheya, A., Bronskill, J. F., Patacchiola, M., Nowozin, S., and Turner, R. E. Fit: Parameter efficient few-shot transfer learning for personalized and federated image classification. In *International Conference on Learning Representations*, 2023.

Suri, A., Zhang, X., and Evans, D. Do parameters reveal more than loss for membership inference? *Transactions on Machine Learning Research*, 2024.

Thaker, P., Setlur, A., Wu, Z. S., and Smith, V. On the benefits of public representations for private transfer learning under distribution shift. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pp. 27088–27120. Curran Associates, Inc., 2024.

Tobaben, M., Shysheya, A., Bronskill, J. F., Paverd, A., Tople, S., Zanella-Beguelin, S., Turner, R. E., and Honkela, A. On the efficacy of differentially private few-shot image classification. *Transactions on Machine Learning Research*, 2023.

Tobaben, M., Ito, H., Jälkö, J., Pradhan, G., He, Y., and Honkela, A. Impact of dataset properties on membership inference vulnerability of deep transfer learning. *arXiv preprint arXiv:2402.06674*, 2024.

Tramèr, F., Kamath, G., and Carlini, N. Position: Considerations for differentially private learning with large-scale public pretraining. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 48453–48467. PMLR, 2024.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 210–218. Springer-Verlag, 2018.

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., and Shokri, R. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, pp. 3093–3106. Association for Computing Machinery, 2022.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE Computer Society, 2018.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 27, pp. 3320–3328. Curran Associates, Inc., 2014.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016*. BMVA Press, 2016.

Zarifzadeh, S., Liu, P., and Shokri, R. Low-cost high-power membership inference attacks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 58244–58282. PMLR, 2024.

Ziyin, L., Liu, K., Mori, T., and Ueda, M. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2022.

A.1. Influence of γ in RMIA Scoring Function

As shown in Equation (2), RMIA uses γ (≥ 1) as the threshold for the likelihood ratio test. It determines how much higher should the likelihood of observing model parameters θ be if a target sample x was in the training dataset relative to a random population sample z to pass the membership test. As such, γ is a critical parameter in the RMIA. Following the same efficient implementation as described in Section 5.3, we conduct a sensitivity analysis varying γ from 1 to 64 and report the MIA efficacy to evaluate attack performance. While RMIA efficacy against models trained from scratch is robust to the choice of γ as shown by Zarifzadeh et al. (2024), this robustness does not extend to few-shot transfer learning setting. For ViT-B/16 models fine-tuned on CIFAR-10 with the Head-only setting, only RMIA with $\gamma = 2$ performs consistently across varying the number of shots. Despite wide error bars, its median performance remains stable.

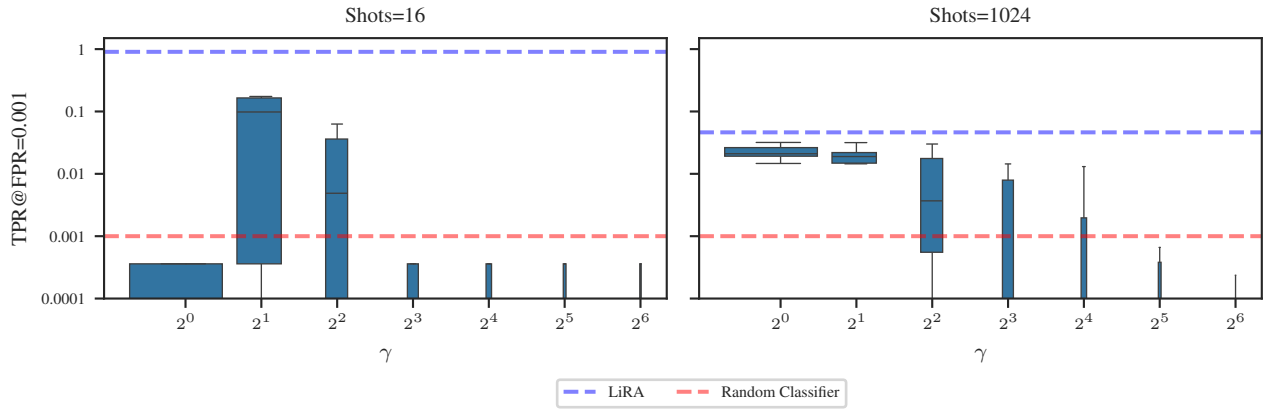

Figure A1. RMIA efficacy as a function of the threshold parameter γ. Results show MIA efficacy against ViT-B/16 models Head-only fine-tuned on CIFAR-10. We train $M + 1$ models ($M = 64$) per repeat, using each model as the target while the remaining M serve as shadow models. We compute the average MIA efficacy (TPR at fixed FPR) across all $M + 1$ target models per repeat, then construct boxplots using the average TPR from 10 independent repeats. RMIA is shown to be sensitive to the value of γ in deep transfer learning setting. The blue dashed line represents the median MIA efficacy achieved by LiRA under identical conditions.

A.2. Effect of Distillation Set Size in Trajectory-MIA

Trajectory-MIA requires an auxiliary dataset for knowledge distillation to simulate the target model's training process. Following Liu et al. (2022) we split the given dataset into target, shadow, and distillation datasets, and we construct our distillation datasets using all data that is **not** a part of the target and shadow datasets.

Figure A2 illustrates that the distillation set size has to be sufficiently large for Trajectory-MIA to work effectively. Smaller distillation sets prevent the student model from learning sufficiently diverse examples to generalize to unseen data. However, $|\mathcal{D}^K| \geq 20000$ does not lead to any significant improvement in Trajectory-MIA efficacy. For CIFAR-10, using $|\mathcal{D}^K| = 20000$ for knowledge distillation proves to be more effective than using all available data ($\sim$50000 samples) in very low-shot regimes like $S = 16$. As the number of shots increases, the performance difference between $|\mathcal{D}^K| = 20000$ and $|\mathcal{D}^K| = \text{All}$ decreases, because the available data for distillation become increasingly similar in both scenarios.

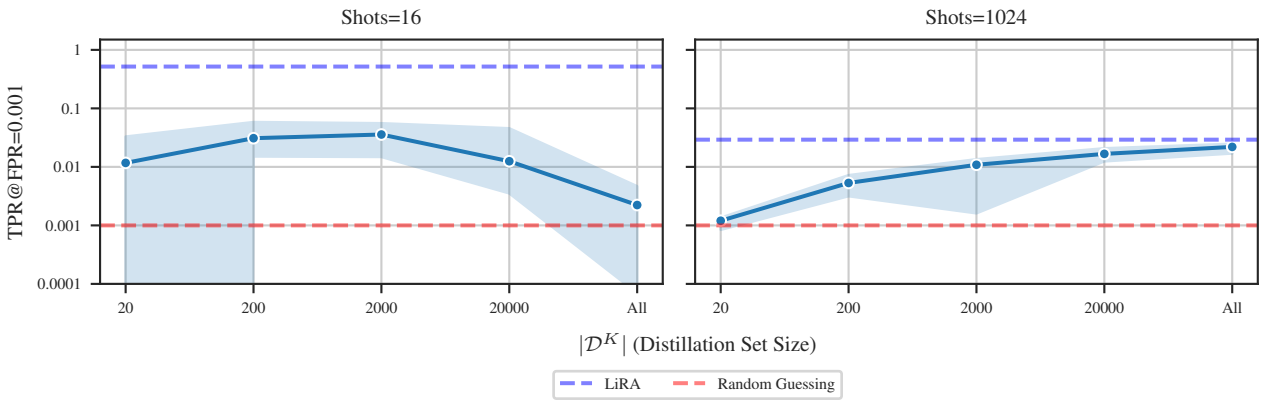

Figure A2. Sensitivity of Trajectory-MIA efficacy to distillation set size. MIA efficacy is evaluated against ViT-B/16 models Head-only fine-tuned on CIFAR-10. $|\mathcal{D}^K|$ represents the distillation set size, with All representing using all available data not part of the fine-tuning datasets. The errorbars represent the interquartile range (IQR) associated with the estimated TPR at fixed FPR. The blue dashed line represents the median MIA efficacy achieved by LiRA under identical conditions. We use LiRA with 64 shadow models as the upper bound on MIA efficacy. Results are averaged over 10 repeats and we use 1 target model per repeat.