[Experiment, Analysis, and Benchmark] Revisiting the LiRA Membership Inference Attack Under Realistic Assumptions

Najeeb Jebreel, Mona Khalil, David Sánchez, and Josep Domingo-Ferrer
Dept. of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Tarragona, Catalonia
{najeeb.jebreel, mona.khalil, david.sanchez, josep.domingo}@urv.cat

Abstract—**Membership inference attacks (MIAs) have become the standard tool for evaluating privacy leakage in machine learning (ML). Among them, the *Likelihood-Ratio Attack* (LiRA) is widely regarded as the state of the art when sufficient shadow models are available. However, prior evaluations have often overstated the effectiveness of LiRA by attacking models overconfident on their training samples, calibrating thresholds on target data, assuming balanced membership priors, and/or overlooking attack reproducibility. We re-evaluate LiRA under a *realistic* protocol that (i) trains models using anti-overfitting (AOF) (and transfer learning (TL), when applicable) to reduce overconfidence as it would be desirable in production models; (ii) calibrates decision thresholds from shadow models and data rather than (usually unavailable) target data; (iii) measures positive predictive value (PPV, a.k.a. precision) under shadow-based thresholds and skewed –rather than unrealistically balanced– membership priors ($\pi \leq 10\%$); and (iv) quantifies per-sample membership reproducibility across different seeds and training variations. In this setting, we find that (a) AOF significantly weakens LiRA and TL further reduces the effectiveness of the attack, while improving model accuracy; (b) with shadow-based thresholds and skewed priors, LiRA's PPV often drops to unreliable levels; and (c) per-sample membership is highly unstable across seeds and training variations even for the strongest LiRA online variant. These results suggest that (i) LiRA, and likely weaker MIAs, are less effective than previously suggested (and often ineffective) in realistic settings; and (ii) for MIAs to serve as meaningful privacy auditing tools, their evaluation must reflect pragmatic training practices, feasible attacker assumptions, and reproducibility considerations. We release our code at: https://github.com/najeebjebreel/lira_analysis.**

Index Terms—**Membership inference attacks, machine learning, privacy, attack reliability, reproducibility.**

I. Introduction

Machine learning (ML) models are frequently trained or fine-tuned on personal data from sensitive domains such as healthcare [1], finance [2], and law [3]. This raises privacy concerns for both individuals and regulators,as trained models can leak information about their training data [4], [5]. Membership inference attacks (MIAs) [6], [7] embody one of these concerns: By inferring whether a specific sample was part of a model training set, an attacker could infer sensitive information (*e.g.*, the participation of an individual in a medical study can indicate that the individual suffers from the disease under study). In this regard, some standardization bodies now classify MIAs as a threat to training data confidentiality [8], [9]. Moreover, due to their simplicity and agnostic nature, MIAs are widely used to empirically assess training data leakage in ML and unlearning [7], [10]–[13].

Why LiRA. The Likelihood-Ratio Attack (LiRA) [14] is widely regarded as the state of the art in MIAs. It formulates membership inference as a hypothesis test using shadow models. Multiple recent studies have shown that LiRA consistently dominates alternatives in the extremely low false positive rates (FPRs) regime [15]–[19], where reliable inference matters [20]. Although newer methods (*e.g.*, Attack R [21], RMIA [22], or LDC-MIA [16]) can achieve a higher true positive rate (TPR) with limited resources, they do not outperform LiRA when sufficient shadow models are available to the attacker (*i.e.*, > 64). Specifically, [18] shows that at an FPR of 0.1%, LiRA identifies substantially more vulnerable samples than RMIA [22] or Attack R [21], and detects nearly all samples labeled as vulnerable by its competitors; [15] show that LiRA is also effective in capturing memorization-based privacy leakage beyond simple overfitting.

We focus our study on LiRA as a privacy benchmark for two reasons: (i) weaker attacks are likely to perform similarly or worse than LiRA; (ii) prior evaluations of LiRA have adopted optimistic choices that may have overestimated its effectiveness. Some of the latter choices are:

1) **Loss-wise overfitting.** Target models typically exhibit a large test-to-train accuracy gap [14], [22]–[24], or a large corresponding *loss ratio* even when the accuracy gap seems to be small (see Table II), reflecting overconfidence in training data that facilitate MIAs [25], [26]. These gaps can be reduced with anti-overfitting techniques [27], [28], which are standard in production settings as they improve model generalization.
2) **Target-based thresholds.** Membership *thresholds* are derived on the (usually unavailable) target model's labeled data. This overfits the decision threshold and unrealistically benefits external attackers.
3) **Balanced priors.** Evaluations assume a balanced mem-

D. Sánchez is an IEEE Senior Member. J. Domingo-Ferrer is an IEEE Fellow. This work was partly supported by the Government of Catalonia (ICREA Acadèmia Prizes to D. Sánchez and J. Domingo-Ferrer, and grant 2021SGR-00115), MCIN/AEI under grant PID2024-157271NB-I00 "CLEARING-IT", and the EU's NextGenerationEU/PRTR via INCIBE (project "HERMES" and INCIBE-URV cybersecurity chair).

bership prior ($\pi = 50\%$), while members, in most cases, are a small set of a larger population [29], [30].

4) **Overlooked reproducibility.** Per-sample reproducibility across different seeds or training variations is often overlooked, leaving open the question of whether attack inferences are stable and reliable.

Moreover, evaluations typically disregard the increasing use of *transfer learning (TL)* in privacy-sensitive domains with limited data [31]–[33], although TL generally improves both model utility and robustness to MIAs compared to training from scratch [19], [34], [35].

This background motivates our four practical questions: **(Q1)** How do AOF and TL affect the utility and effectiveness of LiRA? **(Q2)** What is the effect of shadow-only threshold calibration on attack effectiveness and reliability? **(Q3)** How do the skewed membership priors (*e.g.*, $\pi \leq 10\%$) change PPV at low FPR? **(Q4)** How reproducible is per-sample membership across seeds and training variations?

Contributions. To answer these questions, we reassess LiRA under realistic ML practice and a strong but realistically constrained black-box attacker, with reproducibility in mind.

Specifically, we assume a well-resourced attacker capable of training multiple shadow models on data from the same distribution as the target's training set [7], [14]. We consider the model owner as a pragmatic ML practitioner, who employs anti-overfitting and, when applicable, transfer learning to improve model utility. This setting captures a conservative upper bound on the black-box membership leakage that can be achieved by a strong attacker against well-trained models.

Specifically, our contributions are as follows.

- We design a comprehensive evaluation protocol that (i) systematically varies defender practices (data augmentation, regularization, transfer learning) and attacker assumptions (threshold calibration, membership priors), and (ii) defines consistent evaluation metrics covering *attack effectiveness* (TPR@low-FPR), *reliability* (PPV under realistic priors), and *reproducibility* (inferred samples overlap across runs).
- We show that combining AOF techniques significantly weakens LiRA while preserving model utility, and that TL offer further protection and enhances model utility.
- We show that under shadow-based threshold calibration and realistically skewed priors ($\pi \leq 10\%$), the achieved FPRs deviate from nominal targets, often causing LiRA's PPV to collapse to unreliable levels.
- We quantify reproducibility across random seeds, hyperparameters, and architectures, showing that the sets of "vulnerable" samples identified by LiRA are unstable.
- We identify a strong relationship between the test-to-train loss ratio and LiRA's success, which provides a lightweight, attack-free proxy for monitoring empirical privacy risk. This ratio reflects the global difference in prediction confidence (and, indirectly, uncertainty) between member and non-member data.

Our findings indicate that LiRA (and likely weaker MIAs) are significantly less effective, reliable, and reproducible under realistic conditions than previously suggested. Our evaluation protocol provides concrete guidance for realistic and meaningful privacy audits, emphasizing that MIA evaluations must reflect pragmatic training practices, feasible attacker assumptions, and reproducibility considerations.

II. Background

A. Membership Inference Attacks and LiRA

Membership inference attacks (MIAs) aim to determine whether a sample (x, y) was present in the training set D_{train} of a target model f_θ, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, with $\mathcal{X}$ being the input space and $\mathcal{Y}$ the label space. Since the pioneering work of [7], which introduced the MIA based on "shadow models", several black-box MIAs have been proposed. These methods exploit different signals derived from the model output when queried with a sample, ranging from raw loss or prediction confidence to calibrated, sample-specific scores, or even the final prediction label [14], [21], [22], [36]–[40].

White-box MIAs, which exploit access to model internals, have also been proposed [41], [42] and can sometimes yield marginal gains over black-box methods [42]. However, black-box attacks remain the standard benchmark for assessing membership disclosure and are generally considered more relevant because (i) it is more feasible that attackers can observe only model outputs, and (ii) they are nearly as effective as white-box variants, as the final loss of samples has been shown to be the strongest single predictor of membership [42], [43].

LiRA [14] is widely regarded as the state-of-the-art black-box MIA at extremely low FPRs [16]–[19]. It formulates membership inference as a *statistical hypothesis test* that accounts for *per-sample difficulty* [43]: for a candidate sample (x, y) and target model f_θ, the attacker tests whether $(x, y) \in D_{\text{train}}$ versus $(x, y) \notin D_{\text{train}}$ using shadow models to approximate the distributions of outputs for members versus non-members. Shadow models trained with and without (x, y) produce score distributions that LiRA transforms and models as $\mathcal{N}(\mu_{\text{IN}}, \sigma^2_{\text{IN}})$ and $\mathcal{N}(\mu_{\text{OUT}}, \sigma^2_{\text{OUT}})$. LiRA transforms the sample's true-class confidence $p = f_\theta(x)_y$ using $\phi(p) = \log(p/(1-p))$, which yields approximately Gaussian distributions.

Online LiRA computes the likelihood ratio between the IN and OUT distributions for the observed score $\phi_{\text{obs}} = \phi(f_\theta(x)_y)$, and declares membership when this ratio exceeds a threshold τ chosen to control FPR. *Offline LiRA* reduces computational cost by using only OUT shadow models, evaluating the right-tail probability $p_{\text{out}} = \Pr[Z \geq \phi_{\text{obs}} \mid Z \sim \mathcal{N}(\mu_{\text{OUT}}, \sigma^2_{\text{OUT}})]$ and declaring membership when p_{out} falls below the target FPR. Multivariate distributions using multiple augmentations of the input (*e.g.*, shifts, flips) improve attack effectiveness. Variance estimation can be per sample or fixed globally. With many shadow models (≥ 64), the authors find that the per-sample variances yield stronger results, and the fixed variance is more stable with fewer models.

B. Defenses against MIAs

Defenses against MIAs fall into two broad categories: *formal* methods with certified privacy guarantees, such as differential privacy (DP), and *empirical* approaches, which mitigate leakage in practice but lack formal guarantees, such as standard regularization and data augmentation [44], [45].

Differential privacy (DP [46]). DP in ML is typically implemented via DP-SGD [47], which clips per-sample gradients and injects Gaussian noise during training to bound the influence of any single record. With a meaningful privacy budget (*i.e.*, $\leq 1\varepsilon$ [48]), DP provides formal guarantees by making membership inference statistically unreliable. However, beyond its substantial computational overhead, DP inevitably degrades the utility of models [28], [49], [50], restricting its adoption in real-world deployments [51].

Empirical defenses. Early work has established a connection between overfitting and MIA vulnerability [7], [36]. This motivates the use of standard anti-overfitting (AOF) techniques (such as early stopping [52], weight decay [53], dropout [54], and data augmentation [55]) which have been shown to mitigate leakage by reducing generalization gaps [7], [28], [37], [43], [56]. Empirical studies [28], [56], [57] demonstrate that these techniques (especially when combined) can substantially reduce MIA success rates. Another line of defense limits the information revealed at inference time [7], [58].

A related technique is transfer learning (TL) [59], which adapts models pre-trained on large-scale datasets (*e.g.*, ImageNet, large text corpora) to smaller, task-specific data via fine-tuning. Unlike training from scratch —where limited datasets often cause overfitting and increase susceptibility to MIAs—, TL reuses broad, stable features and typically requires fewer samples and epochs, reducing dataset-specific memorization. TL is now widely adopted in domains where accuracy is critical but data are scarce, such as healthcare [31], [60], making it relevant for realistic MIA evaluation. Empirical work also confirms its protective effect: fine-tuned models are consistently less vulnerable than scratch-trained ones [19], [34], [35]. Specifically, [34] show that pretraining reshapes the privacy–utility landscape, [35] find that combining TL with randomization weakens label-only MIAs, and [19] provide systematic evidence of reduced attack success under TL.

III. Related Work

A growing body of work has raised concerns about the evaluation of MIAs, including LiRA.

a) Dependence on overfitting and poor calibration: The success of MIAs mainly depends on overfitting and miscalibration, which cause models to be overconfident in their training samples. This facilitates the separation between members and non-members [25]–[27]. [26] show theoretically that the advantage of LiRA (and LiRA-style attacks) grows with *model miscalibration* —–when predicted confidence mismatches the true accuracy—–, and decreases with both the data and model uncertainty. [27] demonstrate that when models are trained to achieve high test accuracy and small generalization gaps, all major MIAs, including LiRA, lose most of their effectiveness. AOF techniques have been shown to substantially reduce MIA success with better privacy-utility trade-off than DP [28], [56], [57]. [61] show that modified MixUp augmentation for vision transformers substantially reduces attention-based membership leakage while preserving or improving model utility. In TL settings, score-based MIA success (including LiRA) declines sharply as model generalization improves [19], and even fine-tuned LLMs such as BERT experience reduced attack advantage, especially in terms of TPR at low FPR [62].

b) Reliability and reproducibility: [63] show that naturally trained deep learning (DL) models tend to behave similarly on training and non-training samples, causing MIAs to produce high FPRs. [64] demonstrate that the overconfidence of modern DL models leads score-based MIAs to produce many false positives. Due to several training and attack randomization factors, different attacks (or even different instances of the same attack) often produce highly non-overlapping subsets of "vulnerable" (*i.e.*, member) samples despite similar global metrics [21], [24], [65]. [24] report that six state-of-the-art MIAs, including LiRA, exhibit low consistency at the sample level between runs (Jaccard < 0.4 at 0.1% FPR), with only the deterministic loss-based attack [36] maintaining consistency. Similar behavior was also observed by [21]. This non-reproducibility undermines per-sample positive inferences from single runs.

c) Evaluation assumptions: Assuming balanced membership priors (50/50) can highly inflate the perceived privacy risk, as training members usually represent a small fraction of the overall population (especially in privacy-sensitive domains). Even a modest FPR can generate many false positives and collapse PPV under skewed priors [29], [63]. [29] emphasize the use of PPV under realistically skewed priors for a realistic evaluation of MIA. Moreover, most evaluations overestimate attack success by tuning decision thresholds directly on the scores computed from the target model on its labeled data, giving attackers an unrealistic advantage in choosing the optimal membership threshold. [62] note that shadow models can be used to choose thresholds to achieve a target FPR, but show that such shadow-derived thresholds transfer poorly across datasets and model architectures. [21] emphasize that, to measure genuine leakage rather than artifacts of prior mismatch, the data used to evaluate MIAs should come from a distribution similar to that of the target model's training data. Otherwise, results may over- or under-estimate the true leakage. Finally, LiRA's evaluation protocol assumes a powerful attacker with near perfect auxiliary data and the ability to replicate the target's architecture and training pipeline [7], [23], [37]. While defining worst-case bounds, this overstates realistic leakage: even mild shadow-target mismatches can affect score distributions and inflate FPR [26], [62], [66].

IV. Evaluation Protocol and Experimental Setup

Although previous work has highlighted MIA (and LiRA) limitations, it has addressed them in isolation. We evaluate LiRA under *realistic joint* conditions by (i) training less overfitted models with lower confidence certainty and maintained accuracy; (ii) calibrating decision thresholds using only shadow models and data; (iii) evaluating PPV under shadow-based thresholds and realistic membership priors ($\pi \leq 10\%$);

and (iv) quantifying *per-sample* reproducibility across runs and training variations. *This integrated evaluation clarifies how methodological choices (training practices, threshold calibration, priors, and reproducibility) affect privacy leakage.*

A. Threat Model

Attacker. Following prior work [7], [14], [22], we assume that the attacker has *black-box* query access to a deployed target model f_θ and aims to infer whether a target sample (x, y) was included in f_θ's training set. Although LiRA entails a high computational cost –since its effective performance requires training and querying hundreds of shadow models– we do not restrict the resources allocated to the attack. We assume a well-resourced attacker capable of training 256 *shadow models*, as considered in the original LiRA paper [14].

The attacker can train shadow models on data drawn from the same distribution as the target's training data, and approximate the target's architecture and hyperparameters [67]–[69]. Notice that the models obtained will inevitably differ due to randomness in the data distribution, stochastic training, or potential small variations in hyperparameters. Importantly, we enforce two realistic constraints: (1) The attacker cannot calibrate the decision threshold from the scores obtained from the target model on its training data. This access would trivialize the attack by directly revealing membership of all training data, which is inconsistent with a realistic black-box threat model. (2) The attacker cannot assume balanced membership priors, as real training data are, most of the time, a small fraction of a larger population, especially in sensitive domains [30], [70].

Instead, thresholds are calibrated exclusively on shadow models, with posterior beliefs adjusted via the Bayes rule using realistic priors π [29], [71]. When π is unknown, plausible priors should be considered. This constitutes a *strong yet realistic* attacker: strong enough to approximate an upper bound on black-box risk while respecting realistic constraints.

Defender. The defender is modeled as a pragmatic ML practitioner who aims to deploy accurate models while simultaneously minimizing privacy leakage. Since MIAs primarily exploit overfitting [36], we assume the defender applies standard AOF techniques (data augmentation, weight decay, dropout and/or early stopping), and TL when applicable. These techniques offer not only favorable privacy–utility trade-offs [28], but also improve the generalizability of the model, which is desirable in production settings. Thus, we assume that the defender has an incentive to rely on tuned combinations of these techniques, which can be achieved through automated hyperparameter optimization [72], [73].

B. Datasets and Models

We evaluated the attack on four datasets. *CIFAR-10/100* [74] each contain 60,000 images of size 32 × 32 with 10/100 classes. *GTSRB* [75] consists of 51,839 traffic sign images (43 classes). *Purchase-100* [7] includes 197,324 purchase records, each with 600 features and 100 classes.

The CIFAR-10/100 and Purchase-100 datasets were used in the LiRA paper [14], while GTSRB was examined in [23]. We focus on the CIFAR datasets because, as shown in [14], CIFAR-10 exhibited vulnerability despite a small train-test accuracy gap, and CIFAR-100 showed the highest leakage among the image datasets. On the other hand, the Purchase-100 dataset is widely adopted in MIA research, but [14] argue that its simplicity makes it a poor proxy for realistic privacy risk assessment; we include it here for completeness. GTSRB serves as a realistic benchmark as it achieves near-perfect accuracy with a minimal train–test loss ratio, making it representative of well-calibrated, high-utility models likely to be deployed in practice.

All input features were normalized using dataset-specific statistics. We used ResNet-18 [76] for CIFAR-10 and GTSRB, WideResNet [77] for CIFAR-100, and a fully connected network (FCN) [14] for Purchase-100, all trained from scratch. Furthermore, we used EfficientNet-V2 [78] for transfer learning experiments.

C. Training Configurations

We define three benchmark configurations with increasing regularization strength:

Baseline. This benchmark replicates the original setup of LiRA [14]: image models trained using three image augmentations (horizontal flip, random crop with reflection padding, Cutout [79]), no dropout, and weight decay $5{\times}10^{-4}$. They were trained from scratch with SGD with momentum 0.9, cosine scheduling with an initial learning rate 0.1 [80], and a batch size 256. The FCN models used the same settings, but with an initial learning rate of 0.01 and a batch size of 128 without data augmentation. All models were trained from scratch for up to 100 epochs with early stopping after 20 epochs. Utility and LiRA evaluations were performed using the best saved checkpoints based on validation accuracy.

Anti-overfitting (AOF). These benchmarks combine comprehensive AOF: multiple image augmentations (horizontal flip, random crop with reflection padding, rotation, ColorJitter, CutMix [81]), dropout (10% vision, 50% tabular) and/or increased weight decay (10^{-3}). These configurations achieved favorable utility–privacy trade-offs with test-to-train loss ratios below 2.0 while maintaining accuracy close to that of baseline. Table I summarizes the configurations; see Section V-E for complete ablation studies and more details on training configurations.

AOF and Transfer learning (AOF+TL). We fine-tuned EfficientNet-V2-S [78] pretrained on ImageNet [82] for 5 epochs using AdamW with OneCycle scheduling [83], batch size 64–128, dropout 25%, weight decay $5{\times}10^{-2}$, and the data augmentations used above.

D. Attack Configuration

Following the original LiRA setting [14], we trained the $M{=}256$ shadow models for each configuration to obtain the IN/OUT score distributions. We constructed balanced (50/50) member/non-member splits so each model (shadow/target)

TABLE I: Benchmark configurations. FLP=H.Flip, CRP=R.Crop with padding, ROT=Rotation, JTR=ColorJitter, CUT=Cutout, CMX=CutMix, DRP=dropout ratio, WD=weight decay.

Dataset	Benchmark	Architecture	FLP	CRP	ROT	JTR	CUT	CMX	DRP (%)	WD
CIFAR-10	baseline	ResNet-18	✓	✓			✓		0	5e-4
	AOF	ResNet-18	✓	✓	✓	✓		✓	10	1e-3
	TL	EfficientNet-V2	✓	✓	✓	✓		✓	25	5e-2
CIFAR-100	baseline	WideResNet	✓	✓			✓		0	5e-4
	AOF	WideResNet	✓	✓	✓	✓		✓	10	1e-3
	TL	EfficientNet-V2	✓	✓	✓	✓		✓	25	5e-2
GTSRB	baseline	ResNet-18	✓	✓					0	5e-4
	TL	EfficientNet-V2	✓	✓	✓	✓		✓	25	5e-2
Purchase-100	baseline	FCN							0	5e-4
	AOF	FCN							50	1e-3

contains roughly half the dataset as members, and every sample is a member in exactly 128 shadow models. In inference, following [14], we adopted 18 deterministic transformed inputs for images (original, horizontal flip, and eight 2-pixel shifts with their reflections). For tabular data, a single query per sample was used. We evaluated both *online* (IN/OUT) and *offline* (OUT-only) LiRA variants, each with per-sample and fix variance (FV) estimation. We also considered the global-thresholding attack, which computes threshold scores of the target model without shadow training to show how average-case metrics (*e.g.*, Acc/AUC) can be misleading.

E. Threshold Calibration

We distinguish between two calibration strategies:

Optimistic. A per-target threshold $\tau(\alpha)$ is computed directly from the predictions of the target model on its own data to achieve a nominal FPR α. This setting, used by [14] and most existing MIAs, assumes privileged access and therefore represents an upper bound on attack performance from the auditor's perspective.

Realistic. Here, thresholds are estimated exclusively from the scores obtained from the predictions of shadow models on their respective shadow data. As described below, we use each of the M shadow models as the target once, and estimate its membership threshold as the *median* over the remaining $M-1$ shadows, that is, $\tau_{\text{shadow}}(\alpha) = \text{median}\big(\{\tau_i(\alpha)\}_{i\neq\text{target}}\big)$, where $\tau_i(\alpha)$ is the target FPR α on shadow i. The median provides a robust estimate while limiting the influence of outliers.

Fig. 1 shows increased threshold variability between targets of the same run, which increases further with additional runs. This variability is expected to alter the achieved TPR (and, more importantly, increase the achieved FPR) compared to the controlled optimistic setting.

F. Evaluation Metrics

Evaluation mode. [14] use one shadow model as the target, and report the metrics on it. In contrast, we adopt a *leave-one-out* mode, where each of the M shadows serves as the target once, with results aggregated across all targets. This captures variability among potential target models and eliminates the selection bias that single-target evaluation obscures.

Utility metrics. We measure training and test accuracies and the mean cross-entropy loss, and report their averages across target models. We define the *loss ratio* as $\text{LR} = L_{\text{test}}/L_{\text{train}}$, which captures confidence-based generalization gaps that MIAs exploit more directly than the 0/1 accuracy gaps [84].

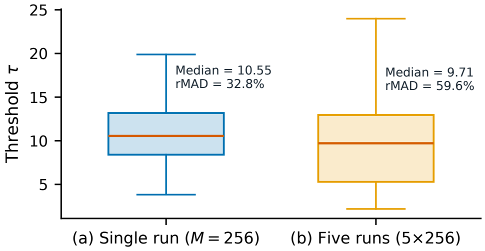


Fig. 1: Distributions of online LiRA thresholds on CIFAR-10 (AOF) at 0.001% FPR. (a) Single run (M=256). (b) Five runs (5×256). Boxplots show the median (τ) and relative median absolute deviation (rMAD). Thresholds exhibit increased variability across targets that widened with multiple runs, reflecting the inherent stochasticity of realistic, target-free calibration.

Privacy metrics. Following [14], [22], we report TPR at extremely low FPRs ($\alpha \in \{0.001\%, 0.1\%\}$) where the inferences are more trustworthy. For a realistic assessment, we compute the predictive positive value (PPV) under skewed priors [29]: $\text{PPV}(\pi) = \pi \cdot \text{TPR'}/(\pi \cdot \text{TPR'} + (1-\pi)\cdot\text{FPR}')$, where FPR$'$ is the achieved (not nominal) FPR, and TPR$'$ is the corresponding achieved TPR. PPV is relevant as it indicates how certain an attacker can be about positive inferences under realistic priors. For these privacy metrics, we report the mean ± standard deviation across target models.

Reproducibility metrics. Following [24], we assess per-sample reproducibility across K independent runs by comparing the sets of samples inferred as members using the threshold corresponding to a target FPR. For each pair of runs, we compute the Jaccard similarity $J(A,B) = |A\cap B|/|A\cup B|$, averaged over all $\binom{K}{2}$ combinations, and also report the average intersection and union sizes of the identified samples. These metrics complement TPR@low-FPR evaluations by quantifying the *consistency* of per-sample inferences under realistic training variability. In addition, we focus on the most vulnerable samples, which are identified as follows: for each run k, we define S_k as the subset of samples flagged as members with *zero per-sample false positives* (0FP) and with within-run support *TP*$\geq x$ (that is, detected by at least x of the 128 IN shadows).

G. Additional Training Details

Baseline vision models followed LiRA [14] with data augmentations consisting of random horizontal flip, reflection-padded 4px crop, and Cutout [79]; no dropout; and weight decay 5×10^{-4}. For AOF benchmarks, we explored multiple augmentation strategies: rotation (±15° for CIFAR-10/100; ±10° for GTSRB) applied with 50% probability; ColorJitter with brightness, contrast, and saturation jitter of ±0.4, and

TABLE II: Model utility and loss ratio across benchmarks.

Benchmark	Train Loss	Test Loss	Loss Ratio	Train Acc (%)	Test Acc (%)
CIFAR-10 (baseline)	0.0032	0.2272	71.0	99.94	93.63
CIFAR-10 (AOF)	0.1351	0.2535	1.88	98.78	94.09
CIFAR-10 (AOF+TL)	0.1210	0.1647	1.36	98.70	97.00
CIFAR-100 (baseline)	0.1198	1.2172	10.16	96.75	69.30
CIFAR-100 (AOF)	0.8007	1.2037	1.50	79.75	68.24
CIFAR-100 (AOF+TL)	0.3373	0.6214	1.84	92.71	83.56
GTSRB (baseline)	0.0448	0.0614	1.37	99.30	98.69
GTSRB (AOF+TL)	0.3145	0.3175	1.01	99.29	98.94
Purchase-100 (baseline)	0.1096	0.1861	1.70	99.70	94.91
Purchase-100 (AOF)	0.2266	0.2763	1.22	96.66	93.18

hue jitter of ±0.1 (parameters reduced by half for GTSRB due to its more constrained color distribution); Cutout [79] (16×16 patches) for baseline configurations; MixUp [85] that linearly interpolates pairs of training samples and their labels with mixing coefficient $\lambda \sim \mathrm{Beta}(1.0, 1.0)$; CutMix [81] that replaces rectangular regions between image pairs while mixing labels proportionally to area, applied with 50% probability and $\lambda \sim \mathrm{Beta}(1.0, 1.0)$. The best AOF combination for vision tasks consists of horizontal flip, reflection-padded crop, rotation, ColorJitter, and CutMix, with 10% dropout and weight decay of 10^{-3} (see paragraph below). For TL, we used the same augmentations but increased dropout to 25% and weight decay to $5{\times}10^{-2}$ to account for the capacity of the pretrained model. For Purchase-100, we compared no dropout with $5{\times}10^{-4}$ weight decay (baseline) against 50% dropout with 10^{-3} weight decay (AOF). No augmentations were applied on tabular data.

Experiments were conducted on Windows 11 Home with an Intel® Core™ i7-12700 (12 cores), 32GB RAM, and an NVIDIA GeForce RTX 4080 (16GB VRAM).

V. Results

A. Impact of Anti-overfitting and Transfer Learning

Model utility. Table II presents model utility metrics across benchmarks. AOF techniques consistently achieved an accuracy comparable to that of baseline models. TL delivered the highest accuracy across all datasets where applied. With CIFAR-100, for example, TL provided an improvement of +14.26% in the test accuracy. We can see that the baseline models exhibited severe loss-wise overfitting with loss ratios reaching 10.16 for CIFAR-100 and 71.0 for CIFAR-10, despite a small test accuracy gap for CIFAR-10. In contrast, AOF and TL dramatically reduced these ratios to below 2, which means that there is no contradiction between utility and reducing vulnerability to MIAs. In our experiments, the loss ratio showed a strong correlation with vulnerability to LiRA (see Section V-D). In particular, the accuracy gaps alone failed to capture the magnitude of the vulnerability because it is insensitive to the confidence distribution. Models can achieve similar test accuracy while exhibiting vastly different loss ratios and corresponding MIA vulnerability. Section V-E presents detailed ablations on individual anti-overfitting components and their optimized combinations.

Attack effectiveness under optimistic evaluation. To isolate the effects of AOF and TL, we first evaluated under target-calibrated thresholds and balanced prior ($\pi = 50\%$), which represents the upper bounds of MIA risk. Tables III–VI present the results. For datasets with high baseline loss ratios (CIFAR-10 and CIFAR-100), all LiRA variants initially achieved high TPR and AUC. In contrast, benchmarks with naturally low loss ratios (GTSRB with 1.37 and Purchase-100 with 1.70) exhibited minimal attack success, especially the well-generalized GTSRB. The introduction of AOF dramatically reduced the attack effectiveness for all benchmarks, especially those with a high baseline loss ratio. For CIFAR-10, the strongest online LiRA variant dropped from 10.268% to 2.723% TPR at FPR=0.1% (a 3.8× reduction) and from 3.956% to 0.248% at FPR=0.001% (a 16× reduction). Adding TL compounded these reductions: the same attack endpoints fell to 0.521% and 0.065%, respectively, representing 20× and 61× reductions from baseline. Similar patterns emerged across all datasets, demonstrating that privacy protection need not compromise (and can actually enhance) model utility. The offline LiRA variants, which rely on one-sided hypothesis testing, suffered even worse once overfitting was controlled and in most cases approached random guessing (AUC≈ 50%). Interestingly, fixed-variance (FV) variants showed marginally better stability than per-sample variance estimation after AOF/TL application, suggesting that per-sample variance estimation becomes unreliable when models generalize well. We can also observe that AUC poorly reflects privacy risk in the ultra-low-FPR regime critical for practical deployment. For example, for the CIFAR-10 baseline, the simple global threshold attack achieved higher AUC than the offline variants, yet yielded TPR far below than the latter. This confirms previous observations that relying on average metrics, such as accuracy or AUC, can be misleading [14], [21].

In summary, across all benchmarks and LiRA variants, the reductions in TPR with AOF ranged from 2.4× to 18× with an average of 6.2×. TL amplified these reductions to 191× with an average of 28×. These dramatic reductions under optimistic evaluation conditions demonstrate that **properly tuned anti-overfitting techniques and transfer learning can cut most of LiRA's effectiveness by addressing the root cause: loss-wise overfitting that creates exploitable confidence disparities between member and non-member samples**.

B. Impact of Skewed Priors and Shadow-based Thresholds

After establishing that AOF and TL substantially reduce the effectiveness of LiRA, we now examine whether any residual success translates into *reliable* risk under *shadow-based* thresholds and *skewed* priors. Tables VII–X report results at a nominal target FPR of 0.001%, contrasting an *optimistic* setting (target-calibrated threshold, π=50%) with a *realistic* setting (shadow-calibrated threshold, $\pi \in \{1\%, 10\%, 50\%\}$). In the optimistic setting, the false positive rate FPR$'$ achieved among all variants was 0, which yielded perfect PPV of 100% for any prior π due to threshold overfitting to the target.

In the realistic setting using *shadow-based* thresholding, the results changed substantially. In the *overfitted baselines* (CIFAR-10/100), PPVs were not markedly affected across pri-

TABLE III: CIFAR-10 with proper AOF and TL. TPR reduction factors relative to baseline shown in parentheses for (AOF) and (TL).

Benchmark	Attack	TPR@0.001% FPR (%)	TPR@0.1% FPR (%)	AUC (%)
Baseline	Online	$3.956_{\pm 1.061}$	$10.268_{\pm 0.555}$	$76.48_{\pm 0.32}$
	Online (FV)	$2.876_{\pm 1.064}$	$9.135_{\pm 0.508}$	$76.28_{\pm 0.31}$
	Offline	$0.762_{\pm 0.348}$	$3.262_{\pm 0.338}$	$55.58_{\pm 0.92}$
	Offline (FV)	$0.948_{\pm 0.526}$	$4.540_{\pm 0.424}$	$56.64_{\pm 0.89}$
	Global	$0.003_{\pm 0.004}$	$0.097_{\pm 0.027}$	$59.97_{\pm 0.32}$
AOF	Online	$0.248_{\pm 0.198}$ (×16)	$2.723_{\pm 0.683}$ (×3.8)	$65.76_{\pm 1.27}$
	Online (FV)	$0.687_{\pm 0.351}$ (×4.2)	$3.483_{\pm 0.405}$ (×2.6)	$68.26_{\pm 1.57}$
	Offline	$0.042_{\pm 0.036}$ (×18)	$0.539_{\pm 0.132}$ (×6.1)	$49.31_{\pm 1.73}$
	Offline (FV)	$0.254_{\pm 0.147}$ (×3.7)	$1.479_{\pm 0.222}$ (×3.1)	$48.99_{\pm 2.35}$
	Global	$0.003_{\pm 0.005}$ (×1.0)	$0.110_{\pm 0.028}$ (×0.9)	$56.19_{\pm 0.50}$
AOF+TL	Online	$0.065_{\pm 0.061}$ (×61)	$0.521_{\pm 0.128}$ (×20)	$56.57_{\pm 0.36}$
	Online (FV)	$0.092_{\pm 0.058}$ (×31)	$0.834_{\pm 0.106}$ (×11)	$57.07_{\pm 0.36}$
	Offline	$0.004_{\pm 0.005}$ (×190)	$0.097_{\pm 0.028}$ (×34)	$49.92_{\pm 1.07}$
	Offline (FV)	$0.016_{\pm 0.020}$ (×59)	$0.253_{\pm 0.079}$ (×18)	$50.03_{\pm 1.22}$
	Global	$0.005_{\pm 0.006}$ (×0.6)	$0.114_{\pm 0.027}$ (×0.9)	$53.06_{\pm 0.26}$

TABLE IV: CIFAR-100 with proper AOF and TL.

Benchmark	Attack	TPR@0.001% FPR (%)	TPR@0.1% FPR (%)	AUC (%)
Baseline	Online	$4.619_{\pm 1.730}$	$15.791_{\pm 0.976}$	$88.28_{\pm 0.21}$
	Online (FV)	$1.730_{\pm 1.158}$	$11.306_{\pm 1.000}$	$87.81_{\pm 0.21}$
	Offline	$0.659_{\pm 0.502}$	$6.044_{\pm 0.640}$	$72.10_{\pm 0.40}$
	Offline (FV)	$0.253_{\pm 0.234}$	$3.911_{\pm 0.524}$	$71.88_{\pm 0.39}$
	Global	$0.002_{\pm 0.004}$	$0.098_{\pm 0.026}$	$69.86_{\pm 0.28}$
AOF	Online	$0.351_{\pm 0.226}$ (×13)	$3.097_{\pm 0.353}$ (×5.1)	$75.32_{\pm 0.31}$
	Online (FV)	$0.214_{\pm 0.162}$ (×8.1)	$2.404_{\pm 0.300}$ (×4.7)	$74.96_{\pm 0.32}$
	Offline	$0.090_{\pm 0.069}$ (×7.3)	$1.141_{\pm 0.205}$ (×5.3)	$58.25_{\pm 0.84}$
	Offline (FV)	$0.105_{\pm 0.086}$ (×2.4)	$1.240_{\pm 0.193}$ (×3.2)	$58.81_{\pm 0.81}$
	Global	$0.004_{\pm 0.006}$ (×0.5)	$0.155_{\pm 0.035}$ (×0.6)	$58.82_{\pm 0.24}$
AOF+TL	Online	$0.270_{\pm 0.237}$ (×17)	$2.367_{\pm 0.398}$ (×6.7)	$67.82_{\pm 0.33}$
	Online (FV)	$0.243_{\pm 0.183}$ (×7.1)	$2.223_{\pm 0.265}$ (×5.1)	$68.07_{\pm 0.38}$
	Offline	$0.012_{\pm 0.013}$ (×55)	$0.294_{\pm 0.087}$ (×21)	$52.56_{\pm 0.98}$
	Offline (FV)	$0.046_{\pm 0.045}$ (×5.5)	$0.734_{\pm 0.179}$ (×5.3)	$53.50_{\pm 1.02}$
	Global	$0.006_{\pm 0.007}$ (×0.3)	$0.158_{\pm 0.035}$ (×0.6)	$58.47_{\pm 0.26}$

TABLE V: GTSRB with proper AOF and TL.

Benchmark	Attack	TPR@0.001% FPR (%)	TPR@0.1% FPR (%)	AUC (%)
Baseline	Online	$0.039_{\pm 0.028}$	$0.274_{\pm 0.059}$	$53.09_{\pm 0.27}$
	Online (FV)	$0.042_{\pm 0.032}$	$0.319_{\pm 0.068}$	$53.10_{\pm 0.28}$
	Offline	$0.002_{\pm 0.004}$	$0.064_{\pm 0.034}$	$49.47_{\pm 0.62}$
	Offline (FV)	$0.005_{\pm 0.007}$	$0.085_{\pm 0.038}$	$49.44_{\pm 0.63}$
	Global	$0.006_{\pm 0.008}$	$0.100_{\pm 0.033}$	$51.10_{\pm 0.29}$
AOF+TL	Online	$0.006_{\pm 0.007}$ (×6.5)	$0.109_{\pm 0.033}$ (×2.5)	$51.68_{\pm 0.40}$
	Online (FV)	$0.019_{\pm 0.017}$ (×2.2)	$0.225_{\pm 0.051}$ (×1.4)	$52.11_{\pm 0.40}$
	Offline	$0.005_{\pm 0.008}$ (×0.4)	$0.094_{\pm 0.036}$ (×0.7)	$49.82_{\pm 0.72}$
	Offline (FV)	$0.006_{\pm 0.008}$ (×0.8)	$0.104_{\pm 0.041}$ (×0.8)	$49.79_{\pm 0.77}$
	Global	$0.007_{\pm 0.009}$ (×0.9)	$0.126_{\pm 0.035}$ (×0.8)	$51.62_{\pm 0.33}$

TABLE VI: Purchase-100 with proper AOF and TL.

Benchmark	Attack	TPR@0.001% FPR (%)	TPR@0.1% FPR (%)	AUC (%)
Baseline	Online	$0.523_{\pm 0.243}$	$4.491_{\pm 0.281}$	$70.16_{\pm 0.29}$
	Online (FV)	$0.180_{\pm 0.110}$	$3.089_{\pm 0.188}$	$69.52_{\pm 0.28}$
	Offline	$0.007_{\pm 0.007}$	$0.500_{\pm 0.077}$	$55.11_{\pm 0.48}$
	Offline (FV)	$0.022_{\pm 0.017}$	$0.713_{\pm 0.078}$	$56.11_{\pm 0.51}$
	Global	$0.001_{\pm 0.001}$	$0.100_{\pm 0.015}$	$54.83_{\pm 0.15}$
AOF	Online	$0.022_{\pm 0.017}$ (×24)	$0.825_{\pm 0.068}$ (×5.4)	$62.64_{\pm 0.16}$
	Online (FV)	$0.026_{\pm 0.019}$ (×6.9)	$0.794_{\pm 0.067}$ (×3.9)	$62.82_{\pm 0.16}$
	Offline	$0.001_{\pm 0.001}$ (×7.0)	$0.139_{\pm 0.022}$ (×3.6)	$51.88_{\pm 0.18}$
	Offline (FV)	$0.003_{\pm 0.003}$ (×7.3)	$0.230_{\pm 0.031}$ (×3.1)	$52.50_{\pm 0.19}$
	Global	$0.001_{\pm 0.002}$ (×1.0)	$0.101_{\pm 0.014}$ (×1.0)	$52.20_{\pm 0.13}$

ors because overfitting dominates the signal. Yet FPR′ became non-zero and exhibited variability. More importantly, once AOF (and especially AOF+TL) were applied, we observed (i) significantly increased and more variable FPR′ and (ii) slight to modest change (mostly decreased) in the achieved TPR′. For example, on CIFAR-10 with AOF, online LiRA's FPR′ increased to $0.033\% \pm 0.466\%$ (std exceeding the mean), which decreased PPV to $90.93\% \pm 12.18\%$ at $\pi{=}10\%$ and to $66.52\% \pm 34.88\%$ at $\pi{=}1\%$. These changes with the corresponding high-variance indicate limited transferability of thresholds when models generalize well. Adding TL further degrades PPV (and increases variance) due to the combined effect of slightly lower TPR′ and higher FPR′. For instance, on CIFAR-10 with AOF+TL, online LiRA's PPV at $\pi{=}10\%$ dropped to $70.75\% \pm 30.40\%$. Less expensive offline variants were even worse and their PPVs frequently fell to unreliable levels across benchmarks under AOF/AOF+TL, particularly for $\pi \leq 10\%$. This is because AOF/TL shrank the separation between IN and OUT logit distributions, causing overlap and, therefore, deviated thresholds from the optimal ones.

Across variants, fixed-variance (FV) LiRA often achieved lower FPR′ and PPV with lower standard deviations than per-sample variance, suggesting that a global variance estimate is more robust once the overfitting is reduced. Regarding priors, PPV's sensitivity was driven primarily by the achieved FPR′: the larger (and more variable) FPR′ was, the stronger the decrease of PPV as π decreased. Models that met deployment standards (*e.g.*, GTSRB) yielded very low attack effectiveness under realistic calibration and priors, with almost no reliable membership signal.

From these observations, two implications emerge:

(1) Poor threshold transferability constrains attack reliability. Even a well-resourced attacker cannot reliably calibrate the thresholds without access to the *target's* score distribution. The thresholds learned from shadows, the only realistic option in black-box settings, deviated from the target FPR, producing high variance among the FPR′, TPR′, and PPV achieved for the targets. This is because AOF/TL shrinks the member vs. non-member confidence distributions, and when combined with training stochasticity, models converge to different local minima with distinct calibration. As a result, likelihood-ratio tails become more model-dependent, causing the locally obtained "optimal" threshold to be *less transferable*. This poor transferability and variance do not render LiRA useless, but they *limit* its utility as a reproducible or reliable attack in realistic black-box scenarios.

(2) Imperfect precision undermines the reliability of individual inferences. Under realistic priors ($\pi \leq 10\%$) and shadow calibration, LiRA's PPV frequently fell well below the near-perfect levels required for confident subject-level claims. With AOF, PPV often decreased to 80–90% in $\pi{=}10\%$ and 60–70% in $\pi{=}1\%$. Adding TL further degraded precision, sometimes to 25–50% in the weakest cases. At these levels, a substantial fraction of flagged samples are false positives, granting individuals considerable *plausible deniability* [86]. This unreliability worsens as the membership prior decreases, which is precisely the relevant regime when targeting specific individuals in sensitive domains.

Together, these findings indicate that **under realistic calibration and priors, LiRA's residual success translates into statistically unreliable inferences rather than actionable privacy leakage.**

TABLE VII: CIFAR-10 under target vs. shadow calibration (target FPR = 0.001%).

Benchmark	Attack	Performance		PPV (%)		
		TPR′ (%)	FPR′ (%)	@π=1%	@π=10%	@π=50%
Target-based thresholds (optimistic)						
Baseline	Online	$3.956_{\pm 1.061}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Online (FV)	$2.876_{\pm 1.064}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Offline	$0.762_{\pm 0.348}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Offline (FV)	$0.948_{\pm 0.526}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
Shadow-based thresholds (realistic)						
Baseline	Online	$3.990_{\pm 0.161}$	$0.002_{\pm 0.003}$	$94.73_{\pm 6.10}$	$99.46_{\pm 0.65}$	$99.94_{\pm 0.07}$
	Online (FV)	$2.912_{\pm 0.142}$	$0.002_{\pm 0.003}$	$93.10_{\pm 8.03}$	$99.26_{\pm 0.91}$	$99.92_{\pm 0.10}$
	Offline	$0.713_{\pm 0.052}$	$0.002_{\pm 0.003}$	$81.31_{\pm 20.20}$	$97.24_{\pm 3.33}$	$99.67_{\pm 0.40}$
	Offline (FV)	$0.918_{\pm 0.068}$	$0.003_{\pm 0.005}$	$81.13_{\pm 21.29}$	$97.03_{\pm 4.06}$	$99.64_{\pm 0.52}$
AOF	Online	$0.224_{\pm 0.482}$	$0.033_{\pm 0.466}$	$66.52_{\pm 34.88}$	$90.93_{\pm 12.18}$	$98.42_{\pm 5.25}$
	Online (FV)	$0.636_{\pm 0.101}$	$0.002_{\pm 0.003}$	$80.13_{\pm 21.52}$	$96.69_{\pm 6.53}$	$99.46_{\pm 2.99}$
	Offline	$0.290_{\pm 4.134}$	$0.262_{\pm 4.138}$	$55.17_{\pm 46.40}$	$73.31_{\pm 28.84}$	$93.37_{\pm 9.32}$
	Offline (FV)	$0.310_{\pm 1.179}$	$0.077_{\pm 1.192}$	$67.96_{\pm 34.53}$	$91.18_{\pm 12.71}$	$98.36_{\pm 6.23}$
AOF+TL	Online	$0.084_{\pm 0.048}$	$0.017_{\pm 0.045}$	$49.13_{\pm 44.90}$	$70.75_{\pm 30.40}$	$91.49_{\pm 12.35}$
	Online (FV)	$0.084_{\pm 0.021}$	$0.002_{\pm 0.003}$	$59.25_{\pm 42.01}$	$83.54_{\pm 18.38}$	$97.22_{\pm 3.42}$
	Offline	$0.027_{\pm 0.085}$	$0.026_{\pm 0.085}$	$32.73_{\pm 46.50}$	$37.30_{\pm 43.64}$	$56.17_{\pm 36.15}$
	Offline (FV)	$0.044_{\pm 0.089}$	$0.033_{\pm 0.089}$	$42.39_{\pm 48.08}$	$52.67_{\pm 40.53}$	$78.43_{\pm 21.99}$

TABLE VIII: CIFAR-100 under target vs. shadow calibration (target FPR = 0.001%).

Benchmark	Attack	Performance		PPV (%)		
		TPR′ (%)	FPR′ (%)	@π=1%	@π=10%	@π=50%
Target-based thresholds (optimistic)						
Baseline	Online	$4.619_{\pm 1.730}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Online (FV)	$1.730_{\pm 1.158}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Offline	$0.659_{\pm 0.502}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Offline (FV)	$0.253_{\pm 0.234}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
Shadow-based thresholds (realistic)						
Baseline	Online	$4.670_{\pm 0.197}$	$0.003_{\pm 0.003}$	$95.19_{\pm 5.67}$	$99.51_{\pm 0.60}$	$99.95_{\pm 0.07}$
	Online (FV)	$1.595_{\pm 0.097}$	$0.002_{\pm 0.003}$	$88.86_{\pm 12.36}$	$98.68_{\pm 1.57}$	$99.85_{\pm 0.18}$
	Offline	$0.555_{\pm 0.057}$	$0.002_{\pm 0.003}$	$78.69_{\pm 22.43}$	$96.64_{\pm 3.96}$	$99.60_{\pm 0.49}$
	Offline (FV)	$0.196_{\pm 0.030}$	$0.003_{\pm 0.003}$	$65.09_{\pm 35.23}$	$90.68_{\pm 10.65}$	$98.71_{\pm 1.63}$
AOF	Online	$0.296_{\pm 0.042}$	$0.002_{\pm 0.003}$	$70.36_{\pm 30.45}$	$93.69_{\pm 7.29}$	$99.19_{\pm 0.98}$
	Online (FV)	$0.170_{\pm 0.027}$	$0.002_{\pm 0.003}$	$65.70_{\pm 35.60}$	$90.71_{\pm 10.78}$	$98.71_{\pm 1.64}$
	Offline	$0.070_{\pm 0.018}$	$0.002_{\pm 0.003}$	$58.18_{\pm 42.73}$	$82.14_{\pm 19.85}$	$96.81_{\pm 4.18}$
	Offline (FV)	$0.076_{\pm 0.019}$	$0.002_{\pm 0.003}$	$59.09_{\pm 42.49}$	$82.72_{\pm 19.71}$	$96.91_{\pm 4.14}$
AOF+TL	Online	$0.240_{\pm 0.038}$	$0.006_{\pm 0.018}$	$64.57_{\pm 35.05}$	$89.52_{\pm 14.82}$	$98.19_{\pm 3.94}$
	Online (FV)	$0.204_{\pm 0.036}$	$0.002_{\pm 0.002}$	$68.38_{\pm 32.86}$	$92.61_{\pm 8.38}$	$99.03_{\pm 1.15}$
	Offline	$0.018_{\pm 0.049}$	$0.011_{\pm 0.047}$	$46.06_{\pm 48.94}$	$53.92_{\pm 42.42}$	$76.41_{\pm 26.52}$
	Offline (FV)	$0.049_{\pm 0.049}$	$0.012_{\pm 0.049}$	$50.43_{\pm 46.40}$	$69.37_{\pm 31.01}$	$91.38_{\pm 11.63}$

TABLE IX: GTSRB under target vs. shadow calibration (target FPR = 0.001%).

Benchmark	Attack	Performance		PPV (%)		
		TPR′ (%)	FPR′ (%)	@π=1%	@π=10%	@π=50%
Target-based thresholds (optimistic)						
Baseline	Online	$0.039_{\pm 0.028}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Online (FV)	$0.042_{\pm 0.032}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Offline	$0.002_{\pm 0.004}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Offline (FV)	$0.005_{\pm 0.007}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
Shadow-based thresholds (realistic)						
Baseline	Online	$0.031_{\pm 0.013}$	$0.003_{\pm 0.005}$	$59.13_{\pm 47.20}$	$71.83_{\pm 33.34}$	$91.61_{\pm 11.37}$
	Online (FV)	$0.035_{\pm 0.013}$	$0.003_{\pm 0.005}$	$57.13_{\pm 47.22}$	$71.27_{\pm 32.66}$	$91.67_{\pm 11.23}$
	Offline	$0.003_{\pm 0.006}$	$0.007_{\pm 0.011}$	$11.57_{\pm 31.58}$	$14.29_{\pm 31.18}$	$24.85_{\pm 34.80}$
	Offline (FV)	$0.004_{\pm 0.005}$	$0.005_{\pm 0.008}$	$35.13_{\pm 47.57}$	$37.59_{\pm 46.02}$	$47.72_{\pm 43.17}$
AOF+TL	Online	$0.038_{\pm 0.109}$	$0.037_{\pm 0.106}$	$26.21_{\pm 43.27}$	$32.50_{\pm 40.06}$	$56.85_{\pm 31.44}$
	Online (FV)	$0.017_{\pm 0.019}$	$0.005_{\pm 0.017}$	$53.83_{\pm 48.80}$	$62.45_{\pm 40.28}$	$83.98_{\pm 19.61}$
	Offline	$0.071_{\pm 0.219}$	$0.069_{\pm 0.215}$	$15.93_{\pm 35.70}$	$22.29_{\pm 33.59}$	$47.10_{\pm 32.04}$
	Offline (FV)	$0.080_{\pm 0.219}$	$0.079_{\pm 0.221}$	$18.00_{\pm 37.39}$	$25.24_{\pm 34.83}$	$51.88_{\pm 30.75}$

TABLE X: Purchase-100 under target vs. shadow calibration (target FPR = 0.001%).

Benchmark	Attack	Performance		PPV (%)		
		TPR′ (%)	FPR′ (%)	@π=1%	@π=10%	@π=50%
Target-based thresholds (optimistic)						
Baseline	Online	$0.523_{\pm 0.243}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Online (FV)	$0.180_{\pm 0.110}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Offline	$0.007_{\pm 0.007}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
	Offline (FV)	$0.022_{\pm 0.017}$	$0.000_{\pm 0.000}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$
Shadow-based thresholds (realistic)						
Baseline	Online	$0.516_{\pm 0.047}$	$0.001_{\pm 0.001}$	$89.93_{\pm 11.15}$	$98.84_{\pm 1.37}$	$99.87_{\pm 0.16}$
	Online (FV)	$0.159_{\pm 0.015}$	$0.001_{\pm 0.001}$	$78.87_{\pm 22.27}$	$96.70_{\pm 3.80}$	$99.61_{\pm 0.47}$
	Offline	$0.004_{\pm 0.002}$	$0.001_{\pm 0.001}$	$55.82_{\pm 48.27}$	$66.20_{\pm 37.65}$	$87.37_{\pm 16.56}$
	Offline (FV)	$0.018_{\pm 0.004}$	$0.001_{\pm 0.001}$	$57.27_{\pm 43.60}$	$80.46_{\pm 21.53}$	$96.32_{\pm 4.79}$
AOF	Online	$0.019_{\pm 0.005}$	$0.001_{\pm 0.001}$	$57.73_{\pm 43.46}$	$81.14_{\pm 20.47}$	$96.63_{\pm 4.01}$
	Online (FV)	$0.022_{\pm 0.005}$	$0.001_{\pm 0.001}$	$58.46_{\pm 42.42}$	$82.88_{\pm 18.74}$	$97.07_{\pm 3.57}$
	Offline	$0.001_{\pm 0.001}$	$0.001_{\pm 0.001}$	$27.29_{\pm 44.29}$	$30.20_{\pm 42.78}$	$42.92_{\pm 40.47}$
	Offline (FV)	$0.002_{\pm 0.001}$	$0.001_{\pm 0.001}$	$44.22_{\pm 48.77}$	$51.95_{\pm 42.69}$	$73.99_{\pm 29.04}$

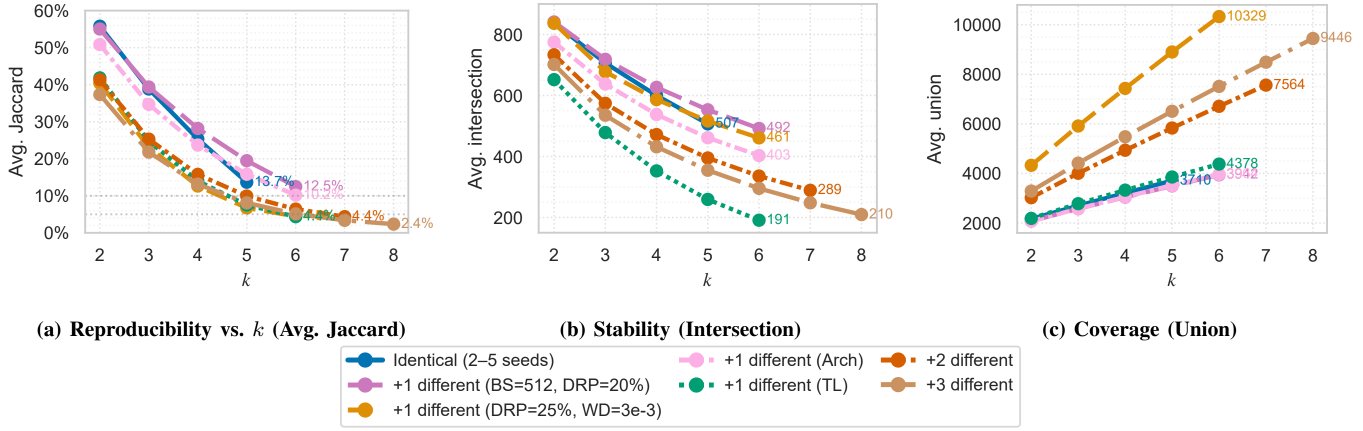
C. Reproducibility Across Seeds and Training Variations

A minimal attack TPR in a realistic setting may still raise concern if the attack achieves high PPV (*e.g.*, online LiRA at TPR@0.001% FPR on realistic CIFAR-10, AOF). However, a critical question remains: *are the identified "vulnerable" samples consistent across runs?* Without reproducibility, the attacker's membership inferences lose credibility and cannot be regarded as reliable evidence of inclusion in training.

We examine whether the online LiRA variant with a strict nominal FPR of 0.001% produces consistent results *in all runs* of CIFAR-10 (AOF). We focus on the online variant because the authors of LiRA [14] consider it as the most effective when the number of shadow models exceeds 64 (we use 256). We ran multiple independent training runs: (i) *identical settings*—-five runs with different seeds but identical hyperparameters; (ii) *minor variations*—-batch size = 512 with dropout = 0.2, and dropout = 0.25 with weight decay = 3×10^{-3}; (iii) *architectural change*—-ResNet18 vs. WideResNet28-2; (iv) *transfer learning*; and (v) *combined variations*.

Severe inconsistency across runs. Fig. 2 shows that, as the number of combined runs increases, the intersection of vulnerable samples (those identified in all runs) shrinks sharply, while the union (samples identified in any run) expands rapidly. This divergence reveals a fundamental instability rather than a consistent identification of genuinely vulnerable samples. Even for *identical settings* with only random seed variation (blue curve), *Jaccard similarity* dropped from about 60% agreement at k=2 runs to less than 14% at k=5. The decay accelerated when additional variations were introduced and combining multiple variations yielded near-zero reproducibility —-an average Jaccard similarity of 2.4% for three distinct configurations (k=8). In other words, more than 97% of the samples flagged as "vulnerable" in one run were not consistently identified in others. This dramatic drop in similarity comes from the decreasing intersection and the rapid expansion of the union, which started at around 2,000 samples and approached 10,000 when all variations were combined. Since the curves show no sign of convergence, it appears that adding more runs would eventually label nearly all training samples as vulnerable. **For an MIA to represent a genuine privacy threat, it must exhibit stable, quasi-deterministic behavior across such variations; otherwise, its per-sample inferences cannot be trusted as reliable indicators of privacy risk.**

Support thresholds and the coverage–stability tradeoff. Fig. 3 analyzes the *zero-FP* detections (the most confident LiRA positives) while varying the support threshold x, *i.e.*, the number of IN shadow models (out of 128) that must agree within a run. A modest increase in support substantially improved stability. At k=5, the average Jaccard similarity rose



(a) **Reproducibility vs. k (Avg. Jaccard)**

(b) **Stability (Intersection)**

(c) **Coverage (Union)**

Fig. 2: Reproducibility, stability, and coverage vs seeds, training variations, and runs (TP≥1).

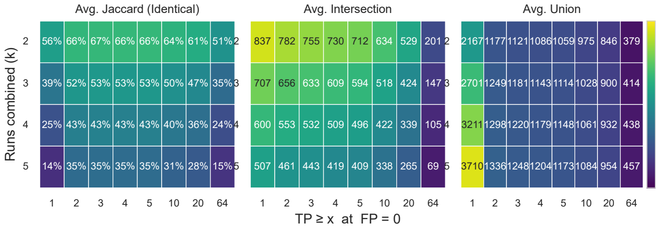


Fig. 3: Reproducibility, stability, and coverage of zero-FP LiRA positives across runs (identical settings). Rows: number of *runs combined* (k); columns: within-run support threshold x (TP$\geq$ x among 128 IN shadows), all at FP$=$ 0. Modest support ($x \in [2, 5]$) improves reproducibility ($\approx$ 35% Jaccard at k=5), while very strong support ($x \geq 20$) reduces both stability and coverage.

from 14% at x=1 to approximately 35% for $x \in [2, 5]$, after which it plateaued. Beyond this range, reproducibility declined again, to 31% at x=10, 28% at x=20, and only 15% at x=64. For any fixed x, the agreement steadily decreased as more runs were combined. These trends confirm considerable stochastic sensitivity, even when the per-sample FPR is zero. Increasing the support threshold also reduced the number of shared detections. At k=5, the intersection shrank from 507 shared positives at x=1 to only 69 at x=64. The union behaved inversely: it expanded with k (as additional runs contributed more single-run detections) but decreased sharply with x, from 3,710 at x=1 to 457 at x=64. Even with FP$=$ 0 per run, most low-support detections were run-specific outliers that failed to generalize across seeds. We therefore recommend support-qualified metrics with $x \in [2, 5]$ at FP$=$ 0 (e.g., TP≥2–5@0FP) as the most meaningful operating region for assessing reproducibility.

Instability of top-ranked vulnerable samples. Even among the 20 most confidently identified vulnerable samples (*i.e.*, those with the highest support within FP$=$ 0 runs), there is little agreement between the runs on the identity or ranking of the samples. New samples frequently appeared among the top-ranked vulnerabilities despite identical training settings. The results of Fig. 3 and Fig. 4 indicate that **"vulnerability" is *not* just an intrinsic property of specific samples, but also influenced by other factors, including model generalization and stochastic training factors such as initialization, mini-batch ordering, and local neighborhood effects [21], [24].**

These reproducibility results do not imply that all samples are immune to attack, but highlight that identifying genuinely vulnerable samples requires repeated experimentation under (at least) varied conditions. Therefore, **when LiRA is applied for privacy auditing, multiple independent runs should be performed to isolate consistently vulnerable samples, which more accurately reflect genuine privacy risk.**

D. Loss Ratio as a Predictor

Across 49 configurations, we observed a clear monotonic relationship between a model's vulnerability to online LiRA and its *loss ratio*. Models with larger loss ratios consistently exhibited higher LiRA success rates, while well-generalized models with ratios below 2 were much less vulnerable. This trend holds across datasets, architectures, and regularization settings, suggesting that the loss ratio could serve as a simple, task-agnostic proxy for privacy risk. It captures in a single quantity how far a model's behavior departs from perfect generalization, linking overfitting directly to the strength of the membership signals available to the attacker.

E. Ablation Study on CIFAR-10

We conducted a comprehensive ablation to isolate the effects of augmentation, dropout, and weight decay on the privacy–utility trade-off. All experiments used WideResNet-28-2 in CIFAR-10 and were evaluated with online LiRA (fixed variance) at 0.1% FPR using target-based thresholds. Table XI reports complete results across 22 configurations.

Augmentation effect. Minimal augmentation (FLP+CRP) led to severe overfitting (loss ratio 713.8) and high vulnerability (TPR 8.47%). Adding Cutout substantially miti-

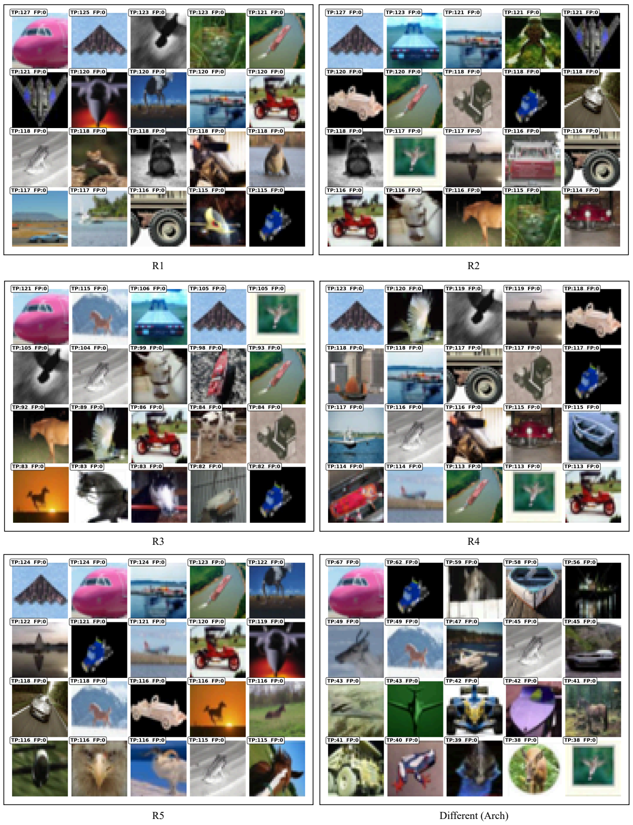


Fig. 4: Top-20 most vulnerable samples across five independent runs (R1–R5; identical settings, different seeds) and one run with a similar architecture (ResNet18–WideResNet28-2).

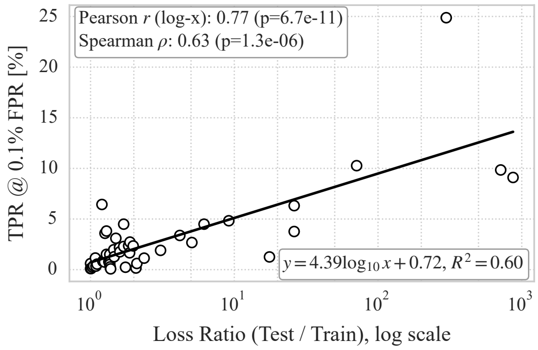


Fig. 5: Correlation between loss ratio (test loss / train loss, log scale) and online LiRA TPR@0.1% FPR.

gated both effects (loss ratio 26.1, TPR 5.93%). Introducing stronger augmentations (rotation, ColorJitter, and CutMix) further improved generalization and privacy (loss ratio 1.35, TPR 0.78%), achieving more than a sevenfold reduction in vulnerability compared to Cutout alone. Across configurations, CutMix consistently outperformed MixUp in privacy protection (TPR 1.87% vs. 1.93%), because its spatially coherent patches discourage pixel-level memorization. Therefore, we selected CutMix for our final configurations.

Dropout effect. Increasing dropout monotonically decreased TPR, though diminishing returns appear beyond 25%. Raising dropout from 25% to 50% reduced TPR by only 1.2% but lowers accuracy by 2.2%. The 25% configuration thus offers an effective balance between privacy (2.24% TPR) and utility (92.2% accuracy).

Weight decay effect. Weight decay exhibited a narrow effective range ($5{\times}10^{-4}$ to 10^{-2}). Lower values ($5{\times}10^{-5}$) provided insufficient regularization (loss ratio 17.6), while excessive decay ($5{\times}10^{-2}$) collapsed the model accuracy to 18.9%. A value of 10^{-3} offered near-optimal stability and privacy.

Combined effects. The best configuration (FLP+CRP+ROT+JTR+CMX, 10% dropout, 10^{-3} weight decay) achieved 1.21% TPR with 93.2% test accuracy, approximately a 5× reduction in attack success compared to the baseline (CRP+FLP+CUT, 0% dropout) while maintaining accuracy. This demonstrates that privacy and utility are not inherently conflicting when anti-overfitting measures are properly combined.

a) Transfer learning and architecture choice: To isolate the impact of transfer learning, we first compared ResNet-18 trained from scratch and with ImageNet pretraining on CIFAR-10. As shown in Table XII, TL reduced LiRA's TPR by more than an order of magnitude while improving loss ratio and accuracy. We further observed that pretrained EfficientNet-V2 yields a comparable drop in attack success but achieves higher test accuracy. Accordingly, we employed EfficientNet-V2 for the main experiments, as a pragmatic practitioner will prefer higher model utility.

VI. Discussion

Overconfidence dominates. Models with high prediction certainty on training data (*i.e.*, loss-wise overfitted) are more vulnerable to attack, regardless of whether shadow-based threshold or skewed priors are applied. Their per-sample losses are often polarized and similar across runs. This observation aligns with empirical findings in [25], which show that the separation between member and non-member increases with model increased confidence on members. However, evaluating LiRA on such overconfident models deviates from realistic deployment and substantially exaggerates its actual privacy threat.

Collapse under realistic conditions. When evaluated under realistic conditions (models trained with AOF or TL to reduce prediction gaps, thresholds calibrated from shadow models, and skewed priors, that is, $\pi \le 10\%$), the apparent success of LiRA vanishes. Membership predictions become statistically unreliable and poorly reproducible because AOF and TL compress the confidence distributions of members and non-members, which LiRA models as Gaussians. As these distributions get closer to each other, the likelihood that the signal of a target model originates from an *OUT* sample increases, pushing thresholds to extreme values and

TABLE XI: Complete ablation study: WideResNet-28-2 on CIFAR-10 with online LiRA (fixed variance) at 0.1% FPR, target-based threshold. Configurations sorted by TPR (descending) within each technique category.

Augmentation	DRP (%)	WD	Train Loss	Test Loss	Loss Ratio	Test Acc (%)	TPR (%)
Augmentation variations							
CRP+FLP	0	5e-4	0.0004	0.2855	713.75	93.03	8.47
CRP+FLP+ROT	0	5e-4	0.0010	0.2389	238.90	93.38	8.01
CRP+FLP+CUT (LiRA)	0	5e-4	0.0079	0.2061	26.09	93.76	5.93
CRP+FLP+ROT+CUT	0	5e-4	0.0241	0.2206	9.15	93.08	4.20
CRP+FLP+ROT+CUT+JTR+MIX	0	5e-4	0.1221	0.2421	1.98	92.66	1.93
CRP+FLP+ROT+JTR+CMX	0	5e-4	0.1804	0.2864	1.59	93.69	1.87
CRP+FLP+ROT+CUT+JTR+MIX+CMX	0	5e-4	0.1700	0.2695	1.59	92.14	1.19
CRP+FLP+ROT+CUT+JTR+CMX	0	5e-4	0.2381	0.3225	1.35	91.31	0.78
Dropout variations							
CRP+FLP+CUT	10	5e-4	0.0440	0.2228	5.06	92.83	3.21
CRP+FLP+CUT	25	5e-4	0.0777	0.2387	3.07	92.16	2.24
CRP+FLP+CUT	35	5e-4	0.1087	0.2567	2.36	91.50	1.67
CRP+FLP+CUT	50	5e-4	0.1719	0.3009	1.75	89.97	1.01
Weight decay variations							
CRP+FLP+CUT	0	5e-5	0.0146	0.2565	17.57	92.56	4.77
CRP+FLP+CUT	0	5e-3	0.1109	0.2329	2.10	92.21	1.39
CRP+FLP+CUT	0	1e-2	0.2368	0.3276	1.38	89.25	0.54
CRP+FLP+CUT	0	5e-2	2.1562	2.1567	1.00	18.92	0.10
Combined strategies (sorted by TPR descending)							
CRP+FLP+CMX	25	5e-4	0.1828	0.3023	1.65	92.25	1.61
CRP+FLP+ROT+JTR+CMX	10	5e-4	0.2142	0.3094	1.44	93.00	1.37
CRP+FLP+ROT+JTR+CMX	10	1e-3	0.2131	0.3029	1.42	93.18	1.21
CRP+FLP+CMX	50	5e-4	0.2586	0.3615	1.40	89.95	0.93
CRP+FLP+ROT+CUT+JTR+CMX	25	5e-4	0.3169	0.3862	1.22	89.03	0.52
CRP+FLP+ROT+CUT+JTR	25	5e-3	0.3319	0.3915	1.18	86.65	0.33

TABLE XII: CIFAR-10 results under the online LiRA attack. The table isolates the effect of TL. Pretrained EfficientNet-V2 (EN-V2) yields a similar effect on attack as pretrained ResNet-18 (RN-18) while providing higher accuracy.

Benchmark	TPR@0.001% FPR (%)	TPR@0.1% FPR (%)	Loss Ratio	Test Acc (%)
Baseline (RN-18)	$3.956_{\pm 1.061}$	$10.268_{\pm 0.555}$	71.0	93.63
AOF (RN-18)	$0.248_{\pm 0.198}$ (×16)	$2.723_{\pm 0.683}$ (×3.8)	1.88	94.09
AOF+TL (RN-18)	$0.076_{\pm 0.055}$ (×52)	$0.782_{\pm 0.171}$ (×13)	1.24	95.10
AOF+TL (EN-V2)	$0.065_{\pm 0.061}$ (×61)	$0.521_{\pm 0.128}$ (×20)	1.36	97.00

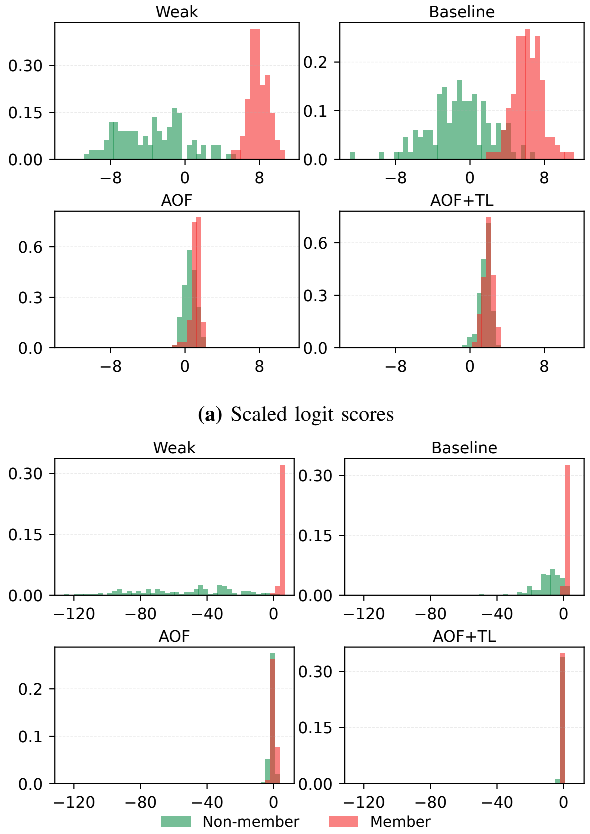
reducing TPR while inflating FPR variance. Fig. 6 illustrates this effect for a representative CIFAR-10 sample. AOF and TL progressively narrow the gap between scaled logit scores, which prevents LiRA from effectively distinguishing between the sample likelihood ratios.

Inference-time augmentation can strengthen the membership signal [14], but its effect fades away when models are not overfitted. Consequently, thresholds estimated from non-overfitted shadow models transfer poorly, producing large gaps between nominal and achieved FPRs. This calibration-induced variance propagates through the Bayes rule into PPV, making per-sample inferences less reliable even when mean TPRs remain moderate. This limitation is not a flaw of LiRA but a statistical inevitability: once member and non-member confidence distributions overlap substantially, no single-sample decision rule can achieve both low FPR and high precision [26].

Poor reproducibility. Across five runs with identical configurations but different seeds, the average Jaccard similarity between detected members was only 13.7%, while the union of flagged examples grew monotonically. This instability indicates that "vulnerability" is not a deterministic property of individual samples, but also depends on the local minimum reached after training, which itself depends on several factors. Thus, each model produces a distinct "vulnerability set" despite nearly identical accuracy and loss. This stochasticity arises from random training and optimization trajectories, amplified by AOF and TL, which induce subtle per-sample confidence shifts. Imposing modest within-run support (TP≥2–5) improves reproducibility but reduces coverage, revealing a clear precision–coverage tension. Claims of persistent per-sample leakage therefore require multi-run confirmation and also clear convergence. When these are absent, most detections likely stem from stochastic artifacts. Even if a few samples are consistently flagged, targeted regularization can address their vulnerability without harming utility [87].

The deployment paradox and privacy–utility synergy. MIAs may constitute genuine privacy threats in domains where training membership is sensitive. However, these domains demand high accuracy and strong regularization, which weaken LiRA. This creates a paradox: *the models most vulnerable to MIAs are those least suitable for deployment in privacy-critical settings.* Our GTSRB results illustrate this. Models that meet deployment standards exhibit natural robustness under realistic thresholds and priors, whereas overfitted baselines exaggerate privacy risk by more than an order of magnitude.

When LiRA may remain relevant and defenses may fail. Our findings do not imply that LiRA is obsolete. Under favorable conditions for attackers (severe overfitting, poor calibration, or unrealistic evaluation setups), LiRA may remain effective. It is a useful *auditing tool* to upper-bound empirical membership leakage. Moreover, practical defenses, such as AOF, TL, and output calibration [58], may not fully



(a) Scaled logit scores

(b) Log-likelihood ratios

Fig. 6: In/out distributions for a representative CIFAR-10 sample under the online attack. AOF and TL significantly narrow the gap between members and non-members.

mitigate attacks when assumptions of data sufficiency and distributional alignment do not hold. Specifically: (i) data scarcity or class imbalance can impair generalization and reintroduce exploitable loss gaps; (ii) domain/temporal shifts between training and non-training data may give attackers an advantage; and (iii) white-box or semi-white-box access (*e.g.*, gradients, internal activations) can amplify leakage; In such cases, empirical auditing remains essential, and combining LiRA-style probing with formal privacy guarantees offers the strongest protection. Differential privacy [46] remains valuable for formal guarantees, but becomes unnecessary when models generalize well enough to suppress empirical leakage.

Need for efficient, reliable and consistent MIAs. [24] proposed ensembling an attack instances (or multiple attacks) across runs to improve stability, but such approaches are computationally prohibitive for large-scale auditing. LiRA already incurs substantial computational overhead. For example, training 256 shadow models on CIFAR-10 with ResNet-18 required approximately 29 hours, representing 256× the cost of training a single target model ($\approx$ 7 minutes). For CIFAR-100 with WideResNet, this increased to 33 hours ($\approx$ 8 minutes per model). Beyond shadow training, LiRA incurred (15-17%) additional overhead for inference and evaluation when 18 inference augmentations were used. This makes larger models intrinsically protected against MIAs. [24] also observed that simple loss-thresholding attacks (the cheapest form of MIA) are often the most consistent across runs, even compared to LiRA. However, their use of a global loss threshold produces a substantial proportion of false positives [14], making such attacks unreliable for per-sample inference.

Recommendations for defenders and evaluators. We recommend: (i) training target models with AOF and/or TL to reduce overconfidence and loss–wise overfitting; (ii) reporting both the loss ratio and training–evaluation loss curves alongside accuracy to expose potential leakage risk; (iii) evaluating attacks under realistic conditions (shadow-calibrated thresholds and skewed priors, $\pi \leq 10\%$) to reflect an external attacker's viewpoint; and (iv) assessing reproducibility across multiple runs to verify the stability of inferred memberships.

Limitations. Our study employs with small- to moderate-scale discriminative models. A similar study on large-scale generative and multimodal systems would require many more computational resources and data. These larger settings, which have shown limited LiRA performance in prior work [14], may exhibit different privacy and generalization behaviors. We also did not consider other realistic conditions, such as significant data distribution shifts or higher training variance, which could further influence attack performance. Moreover, we did not evaluate other MIAs beyond LiRA owing to space limitations, although our evaluation protocol and conclusions are expected to generalize to similar black-box attacks. Finally, loss-ratio values should be interpreted alongside absolute test loss and accuracy, as very low ratios may indicate underfitting rather than genuine robustness.

VII. Conclusions and Future Work

We have revisited LiRA under realistic conditions and found that its membership inference risk has been overstated. When models are properly regularized through anti-overfitting or transfer learning, LiRA's apparent advantage largely disappears without loss of model utility. Shadow-based calibration and realistic, skewed membership priors reveal that LiRA's precision (PPV) collapses for well-generalized models. Furthermore, high disagreement between identical runs exposes poor per-sample reproducibility, challenging LiRA's effectiveness as assessor of "vulnerable" examples.

For practitioners, these findings show that proper use of standard anti-overfitting and/or transfer learning techniques already provides strong empirical privacy protection at no accuracy cost. For researchers and evaluators, our results emphasize the need for overconfidence awareness, shadow-based thresholds, realistic priors, and reproducibility checks when quantifying membership risk.

Future work includes applying our evaluation protocol to larger and more diverse data regimes, including cross-domain and multimodal settings, and exploring adaptive attacks that remain reliable and consistent under realistic evaluation conditions.

AI-Generated Content Acknowledgement

Claude Sonnet 4.5 was used to optimize PyTorch code implementation. ChatGPT-5 was used to correct typos and improve grammar.

References

[1] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[2] P. Gogas and T. Papadimitriou, “Machine learning in economics and finance,” *Computational Economics*, vol. 57, no. 1, pp. 1–4, 2021.

[3] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan, “Chatlaw: Open-source legal large language model with integrated external knowledge bases,” *CoRR*, 2023.

[4] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, “Sok: Security and privacy in machine learning,” in *2018 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2018, pp. 399–414.

[5] M. Rigaki and S. Garcia, “A survey of privacy attacks in machine learning,” *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–34, 2023.

[6] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, “Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays,” *PLoS genetics*, vol. 4, no. 8, p. e1000167, 2008.

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

[8] E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton, “A taxonomy and terminology of adversarial machine learning,” *NIST IR*, vol. 2019, pp. 1–29, 2019.

[9] S. K. Murakonda and R. Shokri, “Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning,” *arXiv e-prints*, pp. arXiv–2007, 2020.

[10] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, “Exploiting unintended feature leakage in collaborative learning,” in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2019, pp. 691–706.

[11] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-Béguelin, “Analyzing leakage of personally identifiable information in language models,” in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 346–363.

[12] M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou, “Towards unbounded machine unlearning,” *Advances in neural information processing systems*, vol. 36, pp. 1957–1987, 2023.

[13] A. Blanco-Justicia, N. Jebreel, B. Manzanares-Salor, D. Sánchez, J. Domingo-Ferrer, G. Collell, and K. Eeik Tan, “Digital forgetting in large language models: A survey of unlearning methods,” *Artificial Intelligence Review*, vol. 58, no. 3, p. 90, 2025.

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

[15] X. Li, Q. Li, Z. Hu, and X. Hu, “On the privacy effect of data enhancement via the lens of memorization,” *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 4686–4699, 2024.

[16] H. Shi, T. Ouyang, and A. Wang, “Learning-based difficulty calibration for enhanced membership inference attacks,” in *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2024, pp. 62–77.

[17] J. Hayes, I. Shumailov, C. A. Choquette-Choo, M. Jagielski, G. Kaissis, K. Lee, M. Nasr, S. Ghalebikesabi, N. Mireshghallah, M. S. M. S. Annamalai *et al.*, “Strong membership inference attacks on massive datasets and (moderately) large language models,” *arXiv preprint arXiv:2505.18773*, 2025.

[18] J. Pollock, I. Shilov, E. Dodd, and Y.-A. de Montjoye, “Free {Record-Level} privacy risk evaluation through {Artifact-Based} methods,” in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 5525–5544.

[19] Y. Bai, G. Pradhan, M. Tobaben, and A. Honkela, “Empirical comparison of membership inference attacks in deep transfer learning,” in *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. [Online]. Available: https://openreview.net/forum?id=eR8mAEt1jT

[20] M. Aerni, J. Zhang, and F. Tramèr, “Evaluations of machine learning privacy defenses are misleading,” in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 1271–1284.

[21] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri, “Enhanced membership inference attacks against machine learning models,” in *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, 2022, pp. 3093–3106.

[22] S. Zarifzadeh, P. Liu, and R. Shokri, “Low-cost high-power membership inference attacks,” in *International Conference on Machine Learning*. PMLR, 2024, pp. 58 244–58 282.

[23] Y. Liu, Z. Zhao, M. Backes, and Y. Zhang, “Membership inference attacks by exploiting loss trajectory,” in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2085–2098.

[24] Z. Wang, C. Zhang, Y. Chen, N. Baracaldo, S. Kadhe, and L. Yu, “Membership inference attacks as privacy tools: Reliability, disparity and ensemble,” *arXiv preprint arXiv:2506.13972*, 2025, to appear at ACM CCS 2025.

[25] Z. Qiao, J. Li, M. Backes, and Y. Zhang, “Overconfidence is a dangerous thing: Mitigating membership inference attacks by enforcing less confident prediction,” in *Proceedings of the 2024 Network and Distributed System Security Symposium (NDSS)*, 2024.

[26] M. Zhu, C. Guo, C. Feng, and O. Simeone, “On the impact of uncertainty and calibration on likelihood-ratio membership inference attacks,” *IEEE Transactions on Information Forensics and Security*, 2025.

[27] A. Dionysiou and E. Athanasopoulos, “Sok: Membership inference is harder than previously thought,” *Proceedings on Privacy Enhancing Technologies*, 2023.

[28] A. Blanco-Justicia, D. Sánchez, J. Domingo-Ferrer, and K. Muralidhar, “A critical review on the use (and misuse) of differential privacy in machine learning,” *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–16, 2022.

[29] B. Jayaraman, L. Wang, K. Knipmeyer, Q. Gu, and D. Evans, “Revisiting membership inference under realistic assumptions,” *Proceedings on Privacy Enhancing Technologies*, vol. 2, pp. 348–368, 2021.

[30] M. B. A. L. . W. S. J. . W. V. A. . M. J. G. . C. M. S. . . Biobank and A. of Us Research Demonstration Project Teams Choi Seung Hoan 14 http://orcid. org/0000-0002-0322-8970 Wang Xin 14 http://orcid. org/0000 0001-6042-4487 Rosenthal Elisabeth A. 15, “Genomic data in the all of us research program,” *Nature*, vol. 627, no. 8003, pp. 340–346, 2024.

[31] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, “Transfer learning for medical image classification: a literature review,” *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022.

[32] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, “Feature transfer learning for face recognition with under-represented data,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5704–5713.

[33] “Transfer learning for financial data predictions: a systematic review,” *arXiv preprint arXiv:2409.17183*, 2024.

[34] X. He and Y. Zhang, “Quantifying and mitigating privacy risks of contrastive learning,” in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 845–863.

[35] A. Rajabi, R. Pimple, A. Janardhanan, S. Asokraj, B. Ramasubramanian, and R. Poovendran, “Double-dip: Thwarting label-only membership inference attacks with transfer learning and randomization,” in *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, 2024, pp. 1937–1939.

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

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

[38] L. Watson, C. Guo, G. Cormode, and A. Sablayrolles, “On the importance of difficulty calibration in membership inference attacks,” in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=3eIrli0TwQ

[39] M. Bertran, S. Tang, A. Roth, M. Kearns, J. H. Morgenstern, and S. Z. Wu, “Scalable membership inference attacks via quantile regression,”

Advances in Neural Information Processing Systems, vol. 36, pp. 314–330, 2023.

[40] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1964–1974.

[41] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[42] T. Leemann, B. Prenkaj, and G. Kasneci, "Is my data safe? predicting instance-level membership inference success for white-box and black-box attacks," in *ICML 2024 Next Generation of AI Safety Workshop*, 2024.

[43] A. Sablayrolles, M. Douze, Y. Ollivier, C. Schmid, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *ICML 2019-36th International Conference on Machine Learning*, vol. 97, 2019, pp. 5558–5567.

[44] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Demystifying membership inference attacks in machine learning as a service," *IEEE Transactions on Services Computing*, vol. 14, no. 06, pp. 2073–2089, 2021.

[45] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[46] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

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

[48] C. Dwork, "A firm foundation for private data analysis," *Communications of the ACM*, vol. 54, no. 1, pp. 86–95, 2011.

[49] Y. Wang, Q. Wang, L. Zhao, and C. Wang, "Differential privacy in deep learning: Privacy and beyond," *Future Generation Computer Systems*, vol. 148, pp. 408–424, 2023.

[50] H. Du, S. Liu, and Y. Cao, "Can differentially private fine-tuning llms protect against privacy attacks?" in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2025, pp. 311–329.

[51] L. D. Vasto-Terrientes, D. Sánchez, and J. Domingo-Ferrer, "Differential privacy in practice: lessons learned from 10 years of real-world applications," *IEEE Security & Privacy*, vol. In press, 2025.

[52] R. Caruana, S. Lawrence, and C. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," *Advances in neural information processing systems*, vol. 13, 2000.

[53] A. Krogh and J. Hertz, "A simple weight decay can improve generalization," *Advances in neural information processing systems*, vol. 4, 1991.

[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[55] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[56] Y. Kaya and T. Dumitras, "When does data augmentation help with membership inference attacks?" in *International conference on machine learning*. PMLR, 2021, pp. 5345–5355.

[57] E. Lomurno and M. Matteucci, "On the utility and protection of optimization with differential privacy and classic regularization techniques," in *International Conference on Machine Learning, Optimization, and Data Science*. Springer, 2022, pp. 223–238.

[58] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 259–274.

[59] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[60] A. Ebbehoj, M. Ø. Thunbo, O. E. Andersen, M. V. Glindtvad, and A. Hulman, "Transfer learning for non-image data in clinical research: a scoping review," *PLOS Digital Health*, vol. 1, no. 2, p. e0000014, 2022.

[61] Q. Zhang, D. Yuan, B. Zhang, B. Yuan, and B. Du, "Membership inference attacks against vision transformers: Mosaic mixup training to the defense," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 1256–1270.

[62] Y. He, B. Li, Y. Wang, M. Yang, J. Wang, H. Hu, and X. Zhao, "Is difficulty calibration all we need? towards more practical membership inference attacks," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 1226–1240.

[63] S. Rezaei and X. Liu, "On the difficulty of membership inference attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7892–7900.

[64] D. Hintersdorf, L. Struppek, and K. Kersting, "To trust or not to trust prediction scores for membership inference attacks," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 3043–3049, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2022/422

[65] J. Zhang, D. Das, G. Kamath, and F. Tramèr, "Position: Membership inference attacks cannot prove that a model was trained on your data," in *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE Computer Society, 2025, pp. 333–345.

[66] W. Fu, H. Wang, C. Gao, G. Liu, Y. Li, and T. Jiang, "Membership inference attacks against fine-tuned large language models via self-prompt calibration," *Advances in Neural Information Processing Systems*, vol. 37, pp. 134 981–135 010, 2024.

[67] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction {APIs}," in *25th USENIX security symposium (USENIX Security 16)*, 2016, pp. 601–618.

[68] O. Bastani, C. Kim, and H. Bastani, "Interpreting blackbox models via model extraction," *arXiv preprint arXiv:1705.08504*, 2017.

[69] B. Wang and N. Z. Gong, "Stealing hyperparameters in machine learning," in *2018 IEEE symposium on security and privacy (SP)*. IEEE, 2018, pp. 36–52.

[70] V. H. Murthy, H. M. Krumholz, and C. P. Gross, "Participation in cancer clinical trials: race-, sex-, and age-based disparities," *Jama*, vol. 291, no. 22, pp. 2720–2726, 2004.

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

[72] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, 2012.

[73] S. Falkner, A. Klein, and F. Hutter, "Bohb: Robust and efficient hyperparameter optimization at scale," in *International conference on machine learning*. PMLR, 2018, pp. 1437–1446.

[74] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[75] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: a multi-class classification competition," in *The 2011 international joint conference on neural networks*. IEEE, 2011, pp. 1453–1460.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[77] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

[78] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[79] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[80] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

[81] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.

[82] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[83] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and*

machine learning for multi-domain operations applications, vol. 11006. SPIE, 2019, pp. 369–386.

[84] J. Li, N. Li, and B. Ribeiro, "Membership inference attacks and defenses in classification models," in *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, 2021, pp. 5–16.

[85] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[86] V. Bindschaedler, R. Shokri, and C. A. Gunter, "Plausible deniability for privacy-preserving data synthesis," *Proceedings of the VLDB Endowment*, vol. 10, no. 5, pp. 481–492, 2017.

[87] J. Li *et al.*, "Mist: Defending against membership inference attacks via membership-invariant subspace training," in *USENIX Security Symposium*, 2024. [Online]. Available: https://www.usenix.org/system/files/usenixsecurity24-li-jiacheng.pdf