Revisiting the LiRA Membership Inference Attack Under Realistic Assumptions

Najeeb Jebreel
Universitat Rovira i Virgili, Catalonia
najeeb.jebreel@urv.cat

Mona Khalil
Universitat Rovira i Virgili, Catalonia
mona.khalil@urv.cat

David Sánchez
Universitat Rovira i Virgili, Catalonia
david.sanchez@urv.cat

Josep Domingo-Ferrer
Universitat Rovira i Virgili, Catalonia
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 from near-perfect to substantially lower levels, especially under AOF/AOF+TL and for $\pi \leq 10\%$; and (c) LiRA's thresholded "vulnerable" sets at extremely low FPR exhibit poor reproducibility across runs, while likelihood ratio-based rankings are more stable. These results suggest that (i) LiRA, and likely weaker MIAs, are less effective than previously suggested, and their positive inferences can be less reliable under 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.

Keywords

machine learning, privacy, membership inference attacks, attack reliability, reproducibility

Proceedings on Privacy Enhancing Technologies 2026(3), 765–786

https://doi.org/10.56553/popets-2026-0105

1 Introduction

Machine learning (ML) models are frequently trained or fine-tuned on personal data from sensitive domains such as healthcare [35], finance [25], and law [15]. This raises privacy concerns for both individuals and regulators, as trained models can leak information about their training data [54, 59]. Membership inference attacks (MIAs) [31, 63] 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 [50, 69]. Moreover, due to their simplicity and agnostic nature, MIAs are widely used to empirically assess training data leakage in ML and unlearning [9, 40, 48, 49, 63].

Why LiRA. The Likelihood-Ratio Attack (LiRA) [11] is widely regarded as the state of the art in MIAs. Multiple recent studies have shown that LiRA consistently dominates alternatives in the extremely low false positive rates (FPRs) regime [4, 26, 44, 55, 62], where reliable inference matters [3]. Although newer methods (*e.g.*, Attack R [78], RMIA [83], or LDC-MIA [62]) 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.*, > 128). Specifically, [55] shows that at an FPR of 0.1%, LiRA identifies substantially more vulnerable samples than RMIA [83] or Attack R [78], and detects nearly all samples labeled as vulnerable by its competitors; [44] show that LiRA is effective in capturing memorization-based leakage beyond simple overfitting.

We focus 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, including:

(1) **Loss-wise overfitting.** Target models typically exhibit a large test-to-train accuracy gap [11, 45, 76, 83], or a large corresponding *loss ratio* even when the accuracy gap seems to be small (see Table 2), reflecting overconfidence in training data that facilitate MIAs [56, 87]. These gaps can be reduced with anti-overfitting (AOF) techniques [10, 18], 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 membership prior ($\pi = 50\%$), while members, in most cases, are a small set of a larger population [8, 33].

(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 [1, 37, 80], although TL generally improves both model utility and robustness to MIAs compared to training from scratch [4, 28, 57].

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 are LiRA outputs across training runs?

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 [11, 63]. 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 across runs* at both levels: (i) thresholded sets inferred at a target FPR and (ii) ranked samples based on granular vulnerability scores (likelihood ratios).
- We show that combining AOF techniques significantly weakens LiRA while preserving model utility, and that TL offers 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 drop from near-perfect (under target-calibrated thresholds) to substantially lower levels. This makes positive inferences less reliable, especially for $\pi \leq 10\%$.
- We quantify reproducibility across training randomness, showing that LiRA's thresholded "vulnerable" sets at extremely low FPR are highly unstable; in contrast, likelihood ratio-based rankings are more stable than thresholded sets, although single-run top-q% identification remains unstable for small q.
- 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.

2 Background

2.1 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 [63], 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 [6, 11, 13, 61, 77–79, 83].

White-box MIAs, which exploit access to model internals, have also been proposed [41, 52] and can sometimes yield marginal gains over black-box methods [41]. However, black-box attacks remain the standard benchmark for assessing membership and are 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 [41, 60].

LiRA [11] is widely regarded as the state-of-the-art black-box MIA at extremely low FPRs [4, 26, 55, 62]. It formulates membership inference as a *statistical hypothesis test* that accounts for *per-sample difficulty* [60]: 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. Beyond thresholded decisions, LiRA's per-sample likelihood ratios provide a continuous vulnerability signal that can support ranking-based analyses across runs. *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 (FV) is more stable with fewer models.

2.2 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 [32, 72].

Differential privacy (DP [20]). DP in ML is typically implemented via DP-SGD [2], 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.*, $\varepsilon \leq 1$ [21]), DP provides formal guarantees by making membership inference statistically unreliable. However, beyond its substantial computational overhead, DP inevitably degrades the utility of models [10, 19, 75], restricting its adoption in real-world deployments [73].

Empirical defenses. Early work has established a connection between overfitting and MIA vulnerability [63, 79]. This motivates the use of standard anti-overfitting techniques (such as early stopping [12], weight decay [39], dropout [67], and data augmentation [14]) which have been shown to mitigate leakage by reducing generalization gaps [10, 36, 60, 61, 63]. Empirical studies [10, 36, 46] demonstrate that these techniques (especially when combined) can substantially reduce MIA success rates. Another line of defense limits the information revealed at inference time [34, 63].

A related technique is transfer learning [53], which adapts models pretrained 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 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 [22, 37], making it relevant for realistic MIA evaluation. Empirical work also confirms its protective effect: fine-tuned models are less vulnerable than scratch-trained ones [4, 28, 57]. Specifically, [28] show that pretraining reshapes the privacy–utility landscape, [57] find that combining TL with randomization weakens label-only MIAs, and [4] provide evidence of reduced attack success under TL.

3 Related Work

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

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 [18, 56, 87]. [87] 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 data and model uncertainty. [18] 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 a better privacy-utility trade-off than DP [10, 36, 46]. [86] 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 [4], and even fine-tuned LLMs such as BERT experience reduced attack advantage, especially in terms of TPR at low FPR [29].

Reliability and reproducibility. [58] show that naturally trained deep learning (DL) models tend to behave similarly on training and non-training samples, causing MIAs to produce high FPRs. [30] 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 [76, 78, 85]. [76] 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 [79] maintaining consistency. Similar behavior was also observed by [78]. This non-reproducibility limits the use of a single run to confidently identify a small, stable subset of "vulnerable" records for targeted mitigation [42].

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 degrade PPV under skewed priors [33, 58]. [33] 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. [29] 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. [78] 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 [45, 61, 63]. While defining worst-case bounds, this overstates realistic leakage: even mild shadow-target mismatches can affect score distributions and inflate FPR [24, 29, 87].

4 Evaluation Protocol and Experimental Setup

Although previous work has highlighted MIA (and LiRA) limitations, it has addressed them in isolation. Such isolated evaluations can miss compounded effects that become apparent only under joint conditions. We evaluate LiRA under *realistic joint* conditions by (i) training less overfitted models with lower prediction-confidence certainty while maintaining 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* variability across runs and training randomness, both for thresholded membership sets and for score/rank stability. *This integrated evaluation clarifies how methodological choices (training practices, threshold calibration, priors, run-to-run variability) affect measured privacy leakage.*

4.1 Threat Model

Attacker. Following prior work [11, 63, 83], 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 [11].

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 [5, 71, 74]. 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 the following 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 [8, 51].

Instead, thresholds are calibrated exclusively on shadow models, with posterior beliefs adjusted via the Bayes' rule using realistic priors π [33, 66]. 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 [79], 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 [10], 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 [23, 65].

4.2 Datasets and Models

We evaluated LiRA on four datasets. *CIFAR-10/100* [38] each contain 60,000 images of size 32×32 with 10/100 classes. *GTSRB* [68] consists of 51,839 traffic sign images (43 classes). *Purchase-100* [63] 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 [11], while GTSRB was examined in [45]. We focus on the CIFAR datasets because, as shown in [11], 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 [11] 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 [27] for CIFAR-10 and GTSRB, WideResNet [82] for CIFAR-100, and a fully connected network (FCN) [11] for Purchase-100, all trained from scratch. Furthermore, we used EfficientNet-V2 [70] for transfer learning experiments.

Table 1: 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

4.3 Training Configurations

We define three benchmarks with increasing regularization strength:

Baseline. This benchmark replicates the original setup of LiRA [11]: image models trained using three image augmentations (horizontal flip, random crop with reflection padding, Cutout [17]), no dropout, and weight decay 5×10^{-4}. They were trained from scratch with SGD with momentum 0.9, cosine scheduling with an initial learning rate 0.1 [47], 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 of no improvement. Utility and LiRA evaluations were performed using the best saved checkpoints based on validation accuracy.

Anti-overfitting. 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 1 summarizes the configurations; see Appendix B for a complete ablation study and additional details on training configurations.

AOF + TL. We fine-tuned ResNet-18 and EfficientNet-V2-S [70] pretrained on ImageNet [16] for 10 and 5 epochs using AdamW with OneCycle scheduling [64], batch size 64–128, dropout 25%, weight decay 5×10^{-2}, and the data augmentations used above. EfficientNet-V2-S offered higher utility while delivering comparable resistance

to LiRA (see Table 15 in Appendix B); thus, we adopted EfficientNet-V2-S for our TL configurations.

Additional training details are presented in Appendix C.

4.4 Attack Configuration

Following the original LiRA setting [11], we trained the M=256 shadow models for each benchmark to obtain the IN/OUT score distributions. We constructed balanced member/non-member splits so each model (shadow/target) contains roughly half of the dataset as members, and every sample is a member in exactly 128 shadow models. In inference, following [11], 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 fixed 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.

4.5 Threshold Calibration

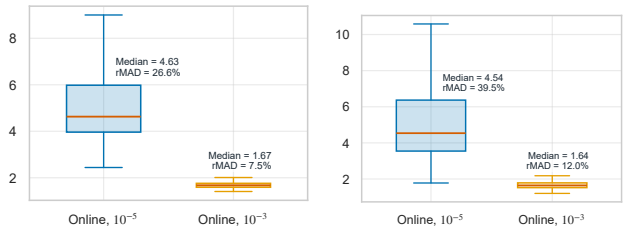
Optimistic. A per-target threshold $\tau(\alpha)$ is computed directly from the predictions of the target model on its own data to achieve an optimal nominal FPR α. This setting, used by [11] and most existing MIAs, assumes privileged access and therefore represents an upper bound on attack performance 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 a larger threshold variability between targets of the same run, which increased further with additional runs. This variability is expected to alter both the achieved TPR and FPR compared to the controlled optimistic setting. It also highlights that pursuing high precision at ultra-low FPR introduces costly instability, whereas relaxing the FPR yields more reproducible thresholds, and this is expected to sacrifice inference precision under realistic skewed membership priors.

4.6 Evaluation Metrics

Evaluation mode. [11] use one shadow model from each run as the target. In contrast, we adopt a *leave-one-out* mode, where each of the M shadows serves as a 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 [43].



(a) Single run **(b) 12 independent runs**

Figure 1: Distributions of decision thresholds of the *online* LiRA variant at nominal FPRs of 0.001% (10^{-5}) and 0.1% (10^{-3}) on CIFAR-10 (ResNet-18, AOF). Boxes show interquartile ranges with medians and relative median absolute deviations (rMAD) from 256 shadow models of a single run (a) and 12 independent runs (b). Threshold variability is substantially higher at 0.001% than at 0.1%, indicating unstable calibration at the extreme tail required for high inference precision.

Privacy metrics. Following [11, 83], 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 positive predictive value (PPV) under skewed priors [33]: $\text{PPV}(\pi) = \pi \cdot \text{TPR'}/(\pi \cdot \text{TPR'} + (1-\pi)\cdot \text{FPR'})$, 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 [76], we assess per-sample reproducibility across K independent runs by comparing the sets of samples inferred as members at a threshold corresponding to a target FPR. Across runs, the training/test split is fixed; variability arises from shadow partitioning, random initialization, and stochastic training dynamics (*e.g.*, batch shuffling). For each pair of runs, we compute the Jaccard similarity $J(A,B) = |A\cap B|/|A\cup B|$, averaged over all $\binom{K}{2}$ pairs, and also report mean $|A\cap B|$ and $|A\cup B|$. To go beyond pairwise comparisons, we additionally consider k-wise agreement over each subset of k runs and average these quantities over all $\binom{K}{k}$ subsets, including $k=K$. In addition, we focus on the most vulnerable samples: for each run r, we define S_r as the subset of samples flagged as members with *zero per-sample false positives* (0FP) and within-run support $TP{\geq}\, x$ (detected by at least x of the 128 IN shadows). We further evaluate score-level stability. For each run r and sample x, we define a vulnerability score using the IN/OUT shadow-model log-likelihood ratios $\Lambda^{(r)}_{\text{IN}}(x)$ and $\Lambda^{(r)}_{\text{OUT}}(x)$ via the median gap $\Delta^{(r)}_{\text{med}}(x) = \text{median}[\log\Lambda^{(r)}_{\text{IN}}(x)] - \text{median}[\log\Lambda^{(r)}_{\text{OUT}}(x)]$. Samples are ranked by $\Delta^{(r)}_{\text{med}}(\cdot)$, and $T^{(r)}_q$ denotes the top-q% set in run r. For each run pair (r,r'), we report set agreement ($|\cap|$, $|\cup|$, Jaccard) and tail rank stability (Spearman on $T^{(r)}_q \cap T^{(r')}_q$), averaged over all $\binom{K}{2}$ pairs; we also report the corresponding all-runs ($k=K$) metrics.

Table 2: 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

5 Results

5.1 Impact of AOF and TL

Model utility. Table 2 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 the vulnerability to LiRA (see Section 5.5). 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 the corresponding MIA vulnerability. Appendix B 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 3–6 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 does not need to compromise (and can actually enhance) the utility of the model. 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$\approx$ 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 the latter. This confirms previous observations that relying on average metrics, such as accuracy or AUC, can be misleading [11, 78].

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**.

Table 3: CIFAR-10 with proper AOF and TL. TPR reduction factors relative to baseline shown in parentheses for AOF and TL. Standard deviations reported in small font.

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 4: 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 5: 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 6: Purchase-100 with proper AOF.

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)	$52.20_{\pm 0.13}$

5.2 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* membership inferences (*i.e.*, high-certainty positives) under *shadow-based* thresholds and *skewed* priors. Tables 7–10 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 priors 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% ± 0.466% (std exceeding the mean), which decreased PPV to 90.93% ± 12.18% at π=10% and to 66.52% ± 34.88% at π=1%. These changes with the corresponding high-variance indicate limited transferability of thresholds when models generalize well. Adding TL further degraded PPV (and increased 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 π=10% dropped to 70.75% ± 30.40%. Less expensive offline variants were worse and their PPVs frequently fell to substantially less reliable 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, deviating thresholds from the optimal ones. PPV's sensitivity to prior was driven primarily by the achieved FPR′: the larger (and more variable) the FPR′, the stronger the decrease in PPV as π decreased. Note that skewed priors do not change the ROC trade-off and are not unique to MIAs, but assuming π=50% can misrepresent MIA risk [33]. Although PPV is prior-dependent, our results show that its degradation was primarily driven by the interaction between skewed priors and miscalibrated FPR′. 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. 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 and TL shrink the member vs. non-member confidence distributions (see Fig. 8), 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*.

(2) Imperfect precision reduces 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 π=10% and 60–70% in π=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* [7] due to the increased uncertainty in positive inferences. This uncertainty also increases as the membership prior decreases, which is the relevant regime when targeting specific individuals in sensitive domains. We do not claim that moderate PPV eliminates privacy risk, especially under skewed priors where it may still provide targets for further investigation. Rather, lower PPV weakens the evidentiary strength of individual membership claims.

Together, these findings indicate that **under realistic calibration and priors, LiRA's residual success often translates into substantially higher uncertainty in positive inferences rather than actionable privacy leakage.**

5.3 Reproducibility

When LiRA is used to select a small subset of training samples for targeted protection or as ground-truth labels for evaluating efficient alternatives [42, 55], a key question is how reproducible the resulting thresholded set is under run-to-run variability. Substantial cross-run variation implies that a single-run thresholded

Table 7: CIFAR-10 under target vs. shadow calibration (target FPR = 0.001%).

Benchmark	Attack	Performance TPR′ (%)	Performance FPR′ (%)	PPV (%) @π=1%	PPV (%) @π=10%	PPV (%) @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 8: CIFAR-100 under target vs. shadow calibration (target FPR = 0.001%).

Benchmark	Attack	Performance TPR′ (%)	Performance FPR′ (%)	PPV (%) @π=1%	PPV (%) @π=10%	PPV (%) @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)	$0.049_{\pm 0.049}$	$0.012_{\pm 0.049}$	$50.43_{\pm 46.40}$	$69.37_{\pm 31.01}$	$91.38_{\pm 11.63}$

set may fail to reliably capture all high-risk samples. We examine whether the online LiRA variant with a strict nominal FPR of 0.001% produces consistent and stable results *in all runs* of CIFAR-10. We focus on the online variant because the authors of LiRA [11] consider it to be most effective when the number of shadow models exceeds 64 (we use 256). We ran multiple independent training runs (fixed training/test split; see Section 4.6): (i) *identical settings*–12 runs with different seeds but identical hyperparameters; (ii) *minor variations*–batch size = 512 with dropout = 0.2, and MixUp augmentation (instead of CutMix) with dropout = 0.15; (iii) *architectural change*–ResNet18 vs. WideResNet28-2; (iv) *transfer learning*; and (v) *combined variations*.

Severe inconsistency across runs. Fig. 2 shows that, as more runs combined, the intersection of vulnerable samples (those identified in all runs) shrank sharply, while the union (samples identified in any run) expanded rapidly. This divergence indicates that the

Table 9: GTSRB under target vs. shadow calibration (target FPR = 0.001%).

Benchmark	Attack	Performance TPR′ (%)	Performance FPR′ (%)	PPV (%) @π=1%	PPV (%) @π=10%	PPV (%) @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 10: Purchase-100 under target vs. shadow calibration (target FPR = 0.001%).

Benchmark	Attack	Performance TPR′ (%)	Performance FPR′ (%)	PPV (%) @π=1%	PPV (%) @π=10%	PPV (%) @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)	$0.002_{\pm 0.001}$	$0.001_{\pm 0.001}$	$44.22_{\pm 48.77}$	$51.95_{\pm 42.69}$	$73.99_{\pm 29.04}$

identity of thresholded "vulnerable" samples is highly sensitive to retraining variability, rather than converging to a small, stable set. Even for *identical settings* with only random seed variation (blue curve), *Jaccard similarity* dropped from about 50% agreement at k=2 runs, to less than 7.6% at k=12. The decay accelerated when additional variations were introduced, and combining multiple variations yielded near-zero reproducibility –an average Jaccard similarity of 2.5% for four distinct configurations (k=16). In other words, more than 97% of the samples flagged as "vulnerable" in any single run were not consistently identified in others. This dramatic drop arises from sensitivity at the target FPR boundary: even if a single model assigns an unusually high confidence to a sample (member or non-member), it can push that sample above the decision threshold and include it in the vulnerable set for that run. This causes near-boundary samples to fluctuate across runs, expanding the union set. Since the union curves show no sign of convergence, it appears that adding more runs would eventually label a large fraction of training samples as vulnerable.

Support thresholds and the coverage–stability tradeoff. Fig. 3 analyzes the *zero-FP* detections while varying the support threshold x, *i.e.*, the number of IN shadow models (out of 128) that must agree within a run. We can see that increasing the support threshold

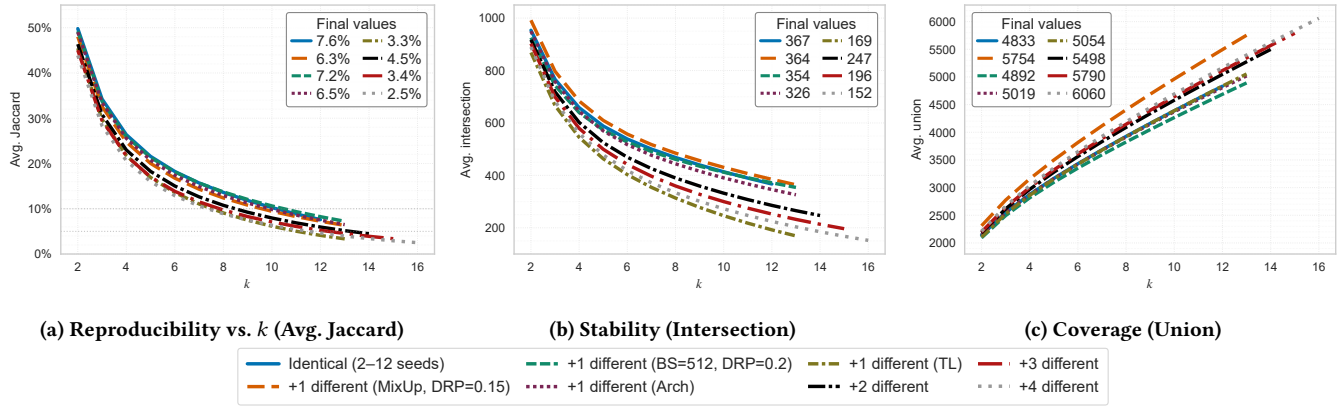


Figure 2: Reproducibility, stability, and coverage vs seeds, training variations, and runs (TP≥1).

yielded only limited gains in reproducibility. The average Jaccard similarity rose slightly from 1.9% at x=1 to a maximum of ≈ 3.2% for $x \in [2, 5]$, after which reproducibility declined again. For any fixed x, agreement decreased sharply as more runs were combined, indicating substantial variability across runs even under FP = 0. Increasing the support threshold also substantially reduced stability and coverage. At k=12, the average intersection shrank from 90 shared positives at x=1 to 0 at x=64. The average union also decreased from 4817 detections at x=1 to 401 at x=64. Even with FP = 0 per run, most low-support detections (TP = 1) are run-specific and fail to generalize across seeds; they reflect edge cases driven by noise and uncertainty in shadow-based thresholds. Support levels in the range $x \in [2, 5]$ substantially reduced the size of the union with only a small to modest decrease in the intersection due to the elimination of noisy detections. Therefore, these support levels represent the most informative range (TP≥ x@0FP) to assess reproducibility.

Instability of top-ranked vulnerable samples. Even among the 9 most confidently identified vulnerable samples (*i.e.*, those with the highest support within FP = 0 runs, see Fig. 4), 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. With more setting variations, disagreement increased more. The results in Fig. 3 and Fig. 4 indicate that **"vulnerability" should *not* be viewed as an intrinsic property of specific samples, but is also influenced by other factors, including model generalization and stochastic training factors such as initialization, mini-batch ordering, and local neighborhood effects [76, 78].** These factors can reshuffle samples across runs, especially those near the extreme low-FPR boundary.

In general, these reproducibility results show that **LiRA's *thresholded* sets at extremely low FPR exhibit limited reproducibility across runs, reducing their reliability for identifying a small, stable set of "vulnerable" samples from a single run.**

Granular score-level validity under run variability. Beyond thresholded sets at a target FPR, we examine whether LiRA's likelihood ratios themselves provide a stable vulnerability ranking signal. For each run, we compute a per-sample score using the `median gap` (Section 4.6). We favor the median over the mean since the `mean gap` exhibits higher variance and weaker extreme-tail stability (see Appendix A). Table 11 reports reproducibility of ranking-based Top-q% sets induced by the `median gap` under two notions of stability: *pairwise* agreement averaged over all pairs of runs (k=2), and *all-runs* agreement requiring membership in all K=12 runs (k=12). We also compare these sets to the corresponding FPR-thresholded selections. The global Spearman correlation over all samples was $83.5\%_{\pm 5.0\%}$, indicating that the overall vulnerability ranking was largely preserved across runs. For small q, pairwise agreement was moderate to high, but agreement across all 12 runs was much stricter. For example, at q=1% ($|T_q|$=500), the average pairwise intersection was 330 ± 51 (Jaccard 50.16% ± 11.27%), while the all-runs intersection contained 134 samples (Jaccard 11.62%). At q=0.1%, where selections were extremely small, all-runs agreement was

Table 11: Reproducibility of ranking-based sets (Top-q% by `median gap`) and FPR-thresholded sets. ρ denotes the Spearman rank correlation, computed on the intersection set. †/‡: Top-q% closely matching the all-runs $|\cap|$ of FPR = 0.001% / 0.1%.

		Pairwise (k = 2 runs)			All-runs (k = 12 runs)		
Top-q%	T_q	Tail ρ (%)	$\lvert\cap\rvert/\lvert\cup\rvert$	Jaccard (%)	Tail ρ (%)	$\lvert\cap\rvert/\lvert\cup\rvert$	Jaccard (%)
0.10	50	$49.45_{\pm 17.95}$	$26/74_{\pm 7}$	$36.02_{\pm 13.33}$	49.44	6/158	3.80
0.25	125	$46.25_{\pm 15.68}$	$72/178_{\pm 16}$	$41.25_{\pm 12.88}$	55.85	21/338	6.21
0.50	250	$47.74_{\pm 16.24}$	$157/343_{\pm 26}$	$46.63_{\pm 11.29}$	46.39	56/621	9.02
0.75	375	$50.58_{\pm 15.30}$	$242/508_{\pm 38}$	$48.58_{\pm 11.23}$	46.77	92/892	10.31
1.00	500	$53.45_{\pm 13.79}$	$330/670_{\pm 51}$	$50.16_{\pm 11.27}$	51.67	134/1153	11.62
2.00	1000	$58.93_{\pm 12.44}$	$721/1279_{\pm 92}$	$57.11_{\pm 10.73}$	57.01	327/2050	15.95
2.24†	1121	$59.70_{\pm 12.56}$	$816/1426_{\pm 102}$	$\mathbf{57.96}_{\pm 10.63}$	57.87	368/**2280**	**16.14**
3.00	1500	$62.16_{\pm 12.43}$	$1123/1877_{\pm 126}$	$60.46_{\pm 10.06}$	62.43	540/2887	18.70
4.00	2000	$64.54_{\pm 12.00}$	$1539/2461_{\pm 161}$	$63.20_{\pm 9.99}$	63.71	807/3675	21.96
5.00	2500	$66.88_{\pm 11.99}$	$1970/3030_{\pm 193}$	$65.63_{\pm 9.90}$	65.88	1103/4392	25.11
10.00	5000	$74.62_{\pm 11.61}$	$4157/5843_{\pm 320}$	$71.61_{\pm 8.82}$	73.17	2633/7851	33.54
10.38‡	5190	$74.97_{\pm 11.58}$	$4327/6053_{\pm 331}$	$\mathbf{71.95}_{\pm 8.84}$	73.36	2749/**8101**	**33.93**
20.00	10000	$81.20_{\pm 10.63}$	$8745/11255_{\pm 460}$	$77.98_{\pm 7.01}$	80.71	6426/13993	45.92
30.00	15000	$84.99_{\pm 9.19}$	$13302/16698_{\pm 503}$	$79.82_{\pm 5.33}$	84.20	10367/20344	50.96
40.00	20000	$87.39_{\pm 7.36}$	$17705/22295_{\pm 573}$	$79.53_{\pm 4.60}$	86.62	13958/27650	50.48
50.00	25000	$88.80_{\pm 5.93}$	$21805/28195_{\pm 884}$	$77.51_{\pm 5.59}$	88.25	16963/37088	45.74
60.00	30000	$89.66_{\pm 4.94}$	$25557/34443_{\pm 1092}$	$74.37_{\pm 5.55}$	89.18	19241/45857	41.96
70.00	35000	$90.10_{\pm 4.18}$	$29411/40589_{\pm 882}$	$72.54_{\pm 3.74}$	89.71	21140/49653	42.58
80.00	40000	$89.79_{\pm 3.74}$	$34382/45618_{\pm 525}$	$75.39_{\pm 2.02}$	90.09	23272/49998	46.55
90.00	45000	$87.75_{\pm 3.97}$	$41147/48853_{\pm 193}$	$84.23_{\pm 0.73}$	90.65	27310/50000	54.62
100.00	50000	$83.46_{\pm 5.00}$	$50000/50000_{\pm 0}$	$100.00_{\pm 0.00}$	83.46	50000/50000	100.00
FPR = 0.001%	–	–	954/2142	49.80	–	367/4833	7.60
FPR = 0.1%	–	–	4239/8454	52.10	–	2748/18132	15.20

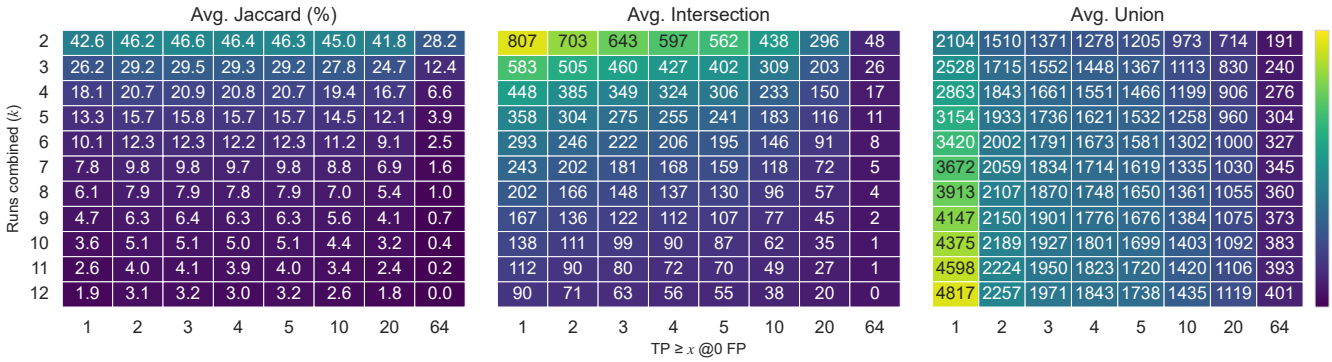


Figure 3: Reproducibility, stability, and coverage of zero-FP LiRA positives across runs (identical settings). Rows: number of *runs combined* (k). Columns: within-run support threshold x (TP$\geq x$ among 128 IN shadows), all at FP$= 0$. Modest support ($x \in [2, 5]$) improves reproducibility, while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

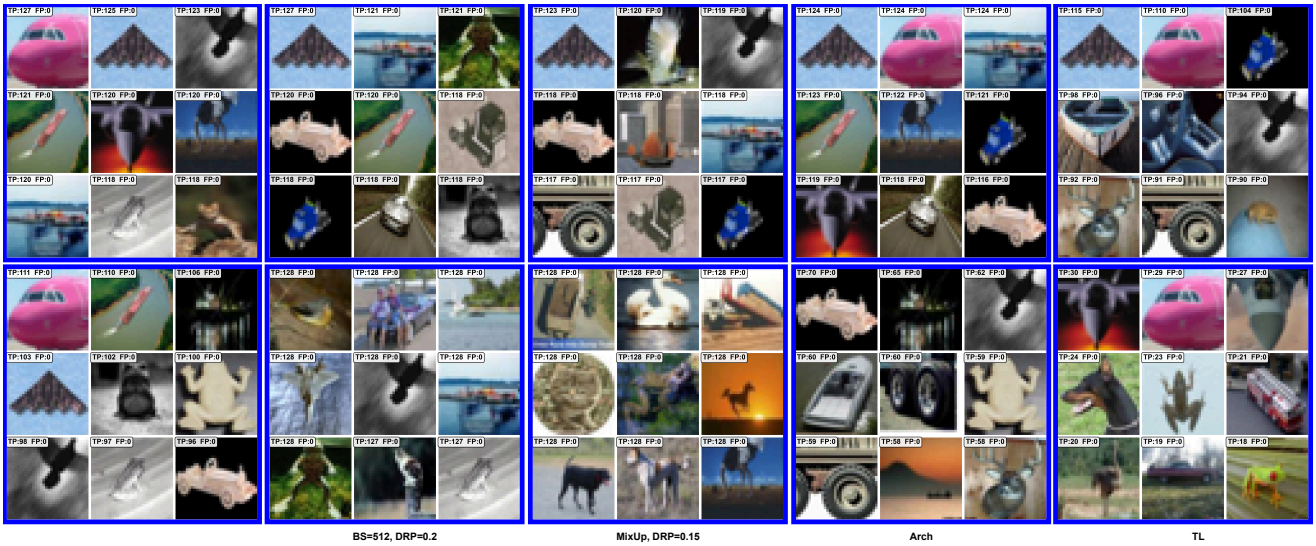


Figure 4: Top-9 most vulnerable samples in 10 different runs. Upper row and leftmost run in the lower row: six independent runs (identical settings, different seeds). Rest of runs in the lower row, respectively: one run with a slight change in batch size and dropout, one run using MixUp instead of CutMix together with a modest dropout increase, one run using a different but similar architecture (ResNet18 vs. WideResNet28-2), and one run with transfer learning. On each image, TP denotes the number of true detections (true positives) among 128 IN shadow models in that run, while FP denotes the number of false detections among the 128 OUT shadow models (FP$= 0$ for all displayed samples).

much stricter (pairwise Jaccard 36.02% ± 13.33% vs. all-runs 3.80%). As q increased, both notions improved (*e.g.*, q=10%: pairwise Jaccard 71.61% ± 8.82% and all-runs 33.54%), consistently with broader high-risk regions being more reproducible. Hence, strict identification in the extreme Top-q% remains the most sensitive regime, and thus a single-run selection may omit samples that rank highly in another run. Disagreement *accumulates* when aggregating across many runs: at q=1%, the average pairwise union was 670 (a +34% expansion over $|T_q|$), while the union over all K=12 runs expanded to 1153 (a +131% expansion), reflecting boundary disagreements that are modest for a pair of run but compound across runs. Figure 5 quantifies how far *displaced* samples fall below the Top-q% cutoff, distinguishing large rank shifts from near-cutoff churn. For each run r, we consider samples that appear in the Top-q% set of at least one run but not in run r, and measure their percentile rank in run r. We then report, for each Δ, the fraction of displaced samples whose percentile rank remains within $q + \Delta$ percentage points. Displacement was predominantly local (especially in the extreme tail): most displaced samples fell just below the cutoff rather than far down the ranking. For instance, at q=1%, 69.73% ± 4.85% of displaced samples remained within q+2% and 89.57% ± 1.98% within q+5%. At the more extreme q=0.1%, 92.52% ± 2.69% remained within q+2%.

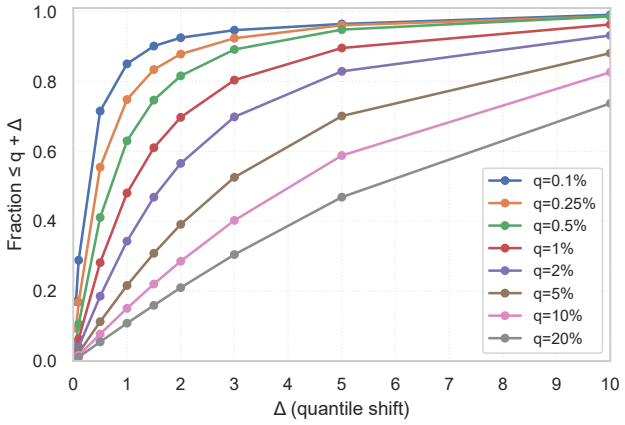


Figure 5: Rank displacement under run-to-run variability. Curves show the fraction of displaced samples whose percentile rank remained within the Top-$(q+\Delta)$%.

This locality helps explaining why unions expand under all-runs aggregation even when cross-run rank changes are mostly near the cutoff. In practice, stability can be improved by selecting a larger Top-q% set in a single run (broader but more reproducible) or by aggregating across runs (*e.g.*, taking the union), at additional computational cost. Figure 13 in Appendix A provides complementary evidence that score dispersion was the largest in the extreme tail.

Matching methods by the strictly stable core (all-runs $|\cap|$, k=12) highlights a clear practical advantage of ranking-based selection. For comparable stable cores, ranking yielded substantially smaller unions and higher all-runs Jaccard than FPR-thresholding: at FPR = 0.001%, thresholding yielded 367/4833 (Jaccard 7.6%), whereas the matched ranking-based choice (q=2.24%) yielded 368/2280 (Jaccard 16.14%); at FPR = 0.1%, thresholding yielded 2748/18132 (Jaccard 15.2%), whereas the matched ranking-based choice (q=10.38%) yielded 2749/8101 (Jaccard 33.93%). Thus, for a fixed stable core size relevant to targeted mitigation, ranking-based selection produced more stable and consistent cross-run sets (smaller union expansion and higher all-runs Jaccard).

Overall, **the likelihood-ratio scores produced by online LiRA retain meaningful rankings that are more stable and consistent across runs than the FPR-thresholded sets, suggesting that LiRA is more reliable as a ranking-based auditing tool than as a precise single-run selector at extreme FPR thresholds**.

5.4 Results with Target FPR of 0.1%

Table 12 shows the results for CIFAR-10 at a target FPR of 0.1% under both optimistic (target-based) and realistic (shadow-based) calibration. As expected, the relaxation of the nominal FPR significantly degraded the PPV for all attack variants for these settings. Fig. 6 presents the corresponding reproducibility analysis. While the more tolerant FPR yielded higher apparent stability than the 0.001% setting, it also substantially increased the size of the union set. Many samples were flagged as vulnerable in at least one run, yet only a small subset was consistently identified across runs. Thus, increasing the target FPR improves agreement at the cost of substantially broader vulnerable sets, reducing selectivity for targeted mitigation. Additional analysis of zero-FP detections under varying within-run support thresholds is presented in Appendix A (Fig. 12).

Table 12: CIFAR-10 under target vs. shadow calibration (target FPR = 0.1%).

Benchmark	Attack	Performance TPR′ (%)	Performance FPR′ (%)	PPV (%) @π=1%	PPV (%) @π=10%	PPV (%) @π=50%
Target-based thresholds (optimistic)						
Baseline	Online	$10.268_{\pm 0.555}$	$0.098_{\pm 0.001}$	$51.271_{\pm 1.378}$	$92.038_{\pm 0.408}$	$99.048_{\pm 0.053}$
	Online (FV)	$9.135_{\pm 0.508}$	$0.098_{\pm 0.001}$	$48.350_{\pm 1.414}$	$91.137_{\pm 0.461}$	$98.931_{\pm 0.060}$
	Offline	$3.262_{\pm 0.338}$	$0.098_{\pm 0.001}$	$25.031_{\pm 1.963}$	$78.499_{\pm 1.803}$	$97.040_{\pm 0.310}$
	Offline (FV)	$4.540_{\pm 0.424}$	$0.098_{\pm 0.001}$	$31.724_{\pm 2.028}$	$83.573_{\pm 1.313}$	$97.860_{\pm 0.202}$
Shadow-based thresholds (realistic)						
Baseline	Online	$10.245_{\pm 0.372}$	$0.101_{\pm 0.022}$	$51.227_{\pm 5.294}$	$91.899_{\pm 1.573}$	$99.027_{\pm 0.204}$
	Online (FV)	$9.151_{\pm 0.347}$	$0.101_{\pm 0.022}$	$48.444_{\pm 5.079}$	$91.036_{\pm 1.657}$	$98.914_{\pm 0.219}$
	Offline	$3.261_{\pm 0.165}$	$0.102_{\pm 0.025}$	$25.271_{\pm 4.670}$	$78.267_{\pm 4.208}$	$96.973_{\pm 0.738}$
	Offline (FV)	$4.520_{\pm 0.213}$	$0.100_{\pm 0.022}$	$32.195_{\pm 5.124}$	$83.561_{\pm 3.106}$	$97.845_{\pm 0.477}$
AOF	Online	$2.826_{\pm 0.750}$	$0.189_{\pm 0.871}$	$21.571_{\pm 7.179}$	$72.813_{\pm 11.409}$	$95.448_{\pm 4.870}$
	Online (FV)	$3.489_{\pm 0.404}$	$0.103_{\pm 0.032}$	$26.554_{\pm 5.090}$	$79.176_{\pm 6.060}$	$97.007_{\pm 2.649}$
	Offline	$0.855_{\pm 5.037}$	$0.434_{\pm 5.077}$	$5.349_{\pm 1.715}$	$37.481_{\pm 7.781}$	$83.563_{\pm 5.673}$
	Offline (FV)	$1.665_{\pm 3.026}$	$0.296_{\pm 3.086}$	$13.403_{\pm 2.987}$	$62.180_{\pm 7.044}$	$93.274_{\pm 4.969}$
AOF+TL	Online	$0.559_{\pm 0.086}$	$0.112_{\pm 0.056}$	$5.244_{\pm 1.278}$	$37.323_{\pm 6.238}$	$83.761_{\pm 4.237}$
	Online (FV)	$0.826_{\pm 0.102}$	$0.101_{\pm 0.025}$	$7.946_{\pm 1.673}$	$48.212_{\pm 5.353}$	$89.161_{\pm 2.022}$
	Offline	$0.126_{\pm 0.128}$	$0.135_{\pm 0.132}$	$0.973_{\pm 0.341}$	$9.669_{\pm 2.850}$	$48.176_{\pm 7.251}$
	Offline (FV)	$0.305_{\pm 0.119}$	$0.133_{\pm 0.115}$	$2.684_{\pm 0.861}$	$22.890_{\pm 5.747}$	$71.679_{\pm 6.875}$

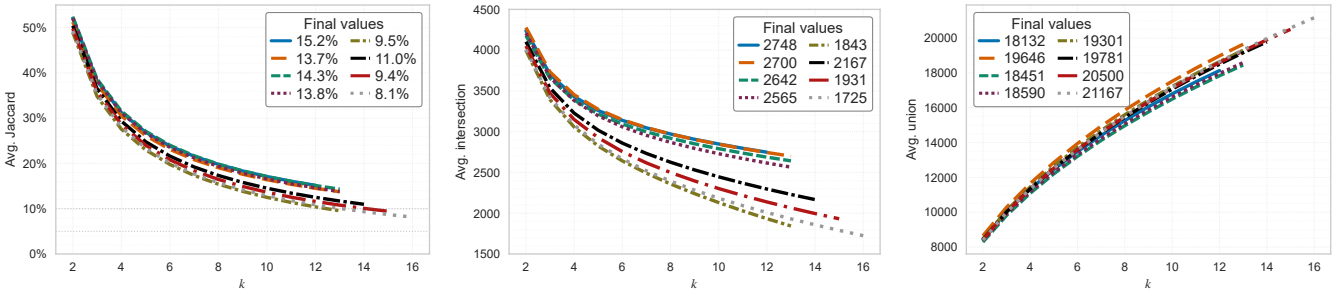
5.5 Loss Ratio as a Predictor

Across 41 configurations, we observed a clear monotonic relationship between a model's vulnerability to online LiRA and its *loss ratio* (see Figure 7). 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.

6 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 the empirical findings in [56], who show that the separation between members and non-members increases with the prediction confidence of the model on members. However, evaluating LiRA on such overconfident models deviates from realistic deployment and substantially exaggerates its actual privacy threat.

Substantial degradation 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, $\pi \leq 10\%$), the apparent success of LiRA substantially degrades. Membership predictions become less reliable and less reproducible because AOF and TL compress the



(a) Reproducibility vs. k (Avg. Jaccard) **(b) Stability (Intersection)** **(c) Coverage (Union)**

Figure 6: Reproducibility, stability, and coverage vs seeds, training variations, and runs (TP≥1 for 0.1%FPR).

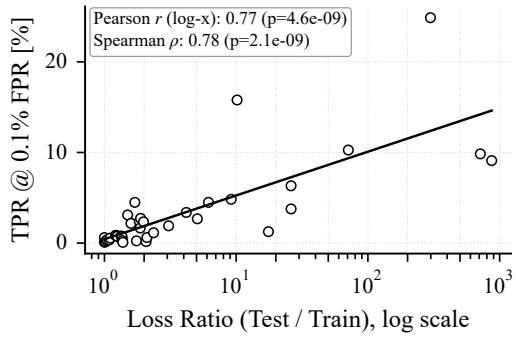
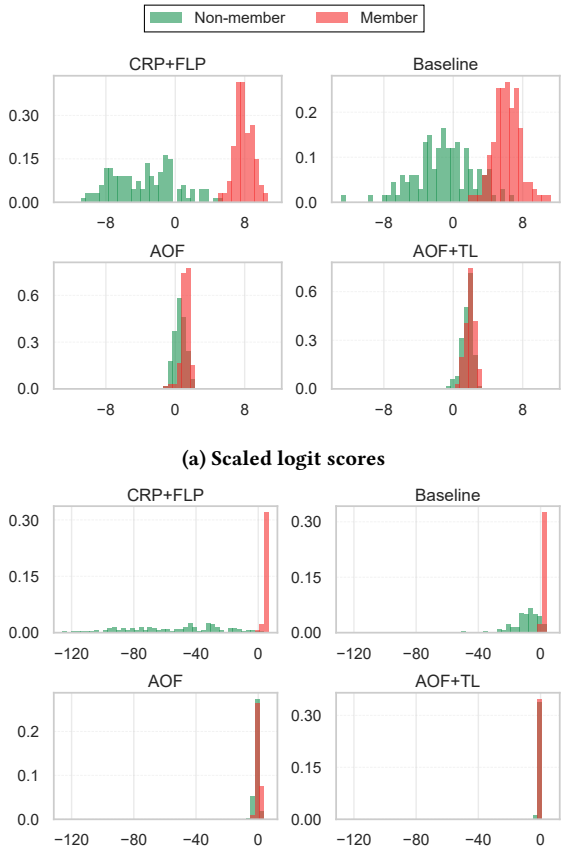


Figure 7: Correlation between loss ratio (test loss / train loss, log scale) and online LiRA TPR@0.1% FPR. Points represent 41 different model configurations



(a) Scaled logit scores

(b) Log-likelihood ratios

Figure 8: 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.

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 reducing TPR while inflating the achieved FPR′ variance. Fig. 8 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 [11], 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 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 [87].

Limited reproducibility. Across 12 runs with identical configurations but different seeds, the average Jaccard similarity between thresholded sets was 7.6% at 0.001% FPR and 15.2% at 0.1% FPR, while the union of flagged samples steadily expanded (especially at 0.1% FPR). This indicates that the *identity of thresholded samples at extreme FPR* is highly sensitive to run-to-run variability, even when aggregate accuracy and loss remain nearly unchanged. In contrast, the underlying LiRA likelihood ratios retain a meaningful and relatively stable *vulnerability ranking* signal. While strict agreement at

small Top-q% selections remains limited (especially under all-runs intersection), most disagreement is concentrated near the cutoff rather than driven by large rank reversals (Figure 5). Practically, these findings suggest that LiRA is better interpreted as a ranking-based auditing tool under retraining variability rather than as a precise single-run selector at extreme thresholds. Stability can be improved by widening the selection (larger Top-q%), increasing within-run support requirements, or aggregating across run-to-run, but each option trades off selectivity and/or computational cost. Importantly, this instability reflects variability *across runs* and does not imply the absence of risk for a fixed deployed model; rather, it limits the reliability of single-run sample-level conclusions.

The deployment paradox and privacy–utility synergy. MIAs may constitute genuine privacy threats in domains where 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 [34], may not fully 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 [20] remains valuable for formal guarantees, but may become unnecessary when models generalize well enough to suppress empirical leakage.

Need for efficient, reliable and consistent MIAs. [76] proposed ensembling several instances of an attack (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, with the equipment mentioned in Appendix C, training 256 shadow models on CIFAR-10 with ResNet-18 took us approximately 29 hours, representing 256× the cost of training a single target model (≈ 7 minutes). For CIFAR-100 with WideResNet, this increased to 33 hours (≈ 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 the deployment of shadow-heavy attacks on large-scale models computationally less practical, increasing the cost for both auditors and potential attackers. [76] also observed that simple loss-thresholding attacks (the cheapest form of MIA) are often the most consistent across runs, even compared to LiRA. However, the use of a global loss threshold produces a substantial proportion of false positives [11], 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 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 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 [11], 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.

Ethical principles. This work did not require the review of an external ethics panel. We evaluated LiRA on publicly available and non-sensitive datasets for realistic privacy assessment, without interacting with human subjects or disclosing sensitive information.

7 Conclusions and Future Work

We reviewed LiRA under realistic conditions and found that its effectiveness for membership inference has been overstated. When models are properly regularized through anti-overfitting or transfer learning, the apparent advantage of LiRA largely decreases without loss of model utility. Shadow-based calibration and realistic, skewed membership priors reveal that LiRA's precision (PPV) drops from near-perfect to substantially lower levels for well-generalized models. Furthermore, we found that *thresholded* vulnerable sets at extreme low-FPR exhibit limited reproducibility across runs. In contrast, LiRA's likelihood-ratio scores retain a meaningful ranking signal, with instability concentrated in the extreme tail. Thus, LiRA is better viewed as a ranking-based auditing tool than as a single-run selector of a small stable subset.

For practitioners, these findings show that proper use of standard anti-overfitting and/or transfer learning techniques can provide 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 will include 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.

Acknowledgments

Partial support to this work has been received from the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and to D. Sánchez), MCIN/AEI under grant PID2024-157271NB-I00 "CLEARING-IT", and the European Commission under project HORIZON-101292277 "SoBigData IP". We used Claude Sonnet 4.5 to optimize code implementation and WriteFull and ChatGPT-5 to correct typos, grammatical errors, and awkward phrasing throughout the article.

References

[1] 2024. Transfer learning for financial data predictions: a systematic review. *arXiv preprint arXiv:2409.17183* (2024).

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[3] Michael Aerni, Jie Zhang, and Florian Tramèr. 2024. Evaluations of machine learning privacy defenses are misleading. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 1271–1284.

[4] Yuxuan Bai, Gauri Pradhan, Marlon Tobaben, and Antti Honkela. 2025. Empirical Comparison of Membership Inference Attacks in Deep Transfer Learning. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*. https://openreview.net/forum?id=eR8mAEt1jT

[5] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504* (2017).

[6] Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. 2023. Scalable Membership Inference Attacks via Quantile Regression. *Advances in Neural Information Processing Systems* 36 (2023), 314–330.

[7] Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. 2017. Plausible deniability for privacy-preserving data synthesis. *Proceedings of the VLDB Endowment* 10, 5 (2017), 481–492.

[8] Mayo Blegen Ashley L. 18 Wirkus Samantha J. 18 Wagner Victoria A. 18 Meyer Jeffrey G. 18 Cicek Mine S. 10 18 Biobank and All 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. 2024. Genomic data in the all of us research program. *Nature* 627, 8003 (2024), 340–346.

[9] Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. 2025. Digital forgetting in large language models: A survey of unlearning methods. *Artificial Intelligence Review* 58, 3 (2025), 90.

[10] Alberto Blanco-Justicia, David Sánchez, Josep Domingo-Ferrer, and Krishnamurty Muralidhar. 2022. A critical review on the use (and misuse) of differential privacy in machine learning. *Comput. Surveys* 55, 8 (2022), 1–16.

[11] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership Inference Attacks From First Principles. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 1897–1914.

[12] Rich Caruana, Steve Lawrence, and C Giles. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in neural information processing systems* 13 (2000).

[13] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-Only Membership Inference Attacks. In *International Conference on Machine Learning*. PMLR, 1964–1974.

[14] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501* (2018).

[15] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR* (2023).

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[17] Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017).

[18] Antreas Dionysiou and Elias Athanasopoulos. 2023. Sok: Membership inference is harder than previously thought. *Proceedings on Privacy Enhancing Technologies* (2023).

[19] Hao Du, Shang Liu, and Yang Cao. 2025. Can Differentially Private Fine-Tuning LLMs Protect Against Privacy Attacks?. In *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 311–329.

[20] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.

[21] Cynthia Dwork. 2011. A firm foundation for private data analysis. *Commun. ACM* 54, 1 (2011), 86–95.

[22] Andreas Ebbehoj, Mette Østergaard Thunbo, Ole Emil Andersen, Michala Vilstrup Glindtvad, and Adam Hulman. 2022. Transfer learning for non-image data in clinical research: a scoping review. *PLOS Digital Health* 1, 2 (2022), e0000014.

[23] Stefan Falkner, Aaron Klein, and Frank Hutter. 2018. BOHB: Robust and efficient hyperparameter optimization at scale. In *International conference on machine learning*. PMLR, 1437–1446.

[24] Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2024. Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems* 37 (2024), 134981–135010.

[25] Periklis Gogas and Theophilos Papadimitriou. 2021. Machine learning in economics and finance. *Computational Economics* 57, 1 (2021), 1–4.

[26] Jamie Hayes, Ilia Shumailov, Christopher A Choquette-Choo, Matthew Jagielski, George Kaissis, Katherine Lee, Milad Nasr, Sahra Ghalebikesabi, Niloofar Mireshghallah, Meenatchi Sundaram Mutu Selva Annamalai, et al. 2025. Strong Membership Inference Attacks on Massive Datasets and (Moderately) Large Language Models. *arXiv preprint arXiv:2505.18773* (2025).

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

[28] Xinlei He and Yang Zhang. 2021. Quantifying and mitigating privacy risks of contrastive learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 845–863.

[29] Yu He, Boheng Li, Yao Wang, Mengda Yang, Juan Wang, Hongxin Hu, and Xingyu Zhao. 2024. Is Difficulty Calibration All We Need? Towards More Practical Membership Inference Attacks. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 1226–1240.

[30] Dominik Hintersdorf, Lukas Struppek, and Kristian Kersting. 2022. 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*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3043–3049. https://doi.org/10.24963/ijcai.2022/422 Main Track.

[31] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics* 4, 8 (2008), e1000167.

[32] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. *ACM Computing Surveys (CSUR)* 54, 11s (2022), 1–37.

[33] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. 2021. Revisiting Membership Inference Under Realistic Assumptions. *Proceedings on Privacy Enhancing Technologies* 2 (2021), 348–368.

[34] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. 2019. MemGuard: Defending Against Black-Box Membership Inference Attacks via Adversarial Examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 259–274.

[35] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* 2, 6 (2020), 305–311.

[36] Yigitcan Kaya and Tudor Dumitras. 2021. When Does Data Augmentation Help With Membership Inference Attacks?. In *International conference on machine learning*. PMLR, 5345–5355.

[37] Hee E Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E Maros, and Thomas Ganslandt. 2022. Transfer learning for medical image classification: a literature review. *BMC medical imaging* 22, 1 (2022), 69.

[38] Alex Krizhevsky. 2009. *Learning Multiple Layers of Features from Tiny Images*. Technical Report. University of Toronto. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[39] Anders Krogh and John Hertz. 1991. A simple weight decay can improve generalization. *Advances in neural information processing systems* 4 (1991).

[40] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2023. Towards unbounded machine unlearning. *Advances in neural information processing systems* 36 (2023), 1957–1987.

[41] Tobias Leemann, Bardh Prenkaj, and Gjergji Kasneci. 2024. 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*.

[42] Jiacheng Li et al. 2024. MIST: Defending Against Membership Inference Attacks via Membership-Invariant Subspace Training. In *USENIX Security Symposium*. https://www.usenix.org/system/files/usenixsecurity24-li-jiacheng.pdf

[43] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. 2021. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*. 5–16.

[44] Xiao Li, Qiongxiu Li, Zhanhao Hu, and Xiaolin Hu. 2024. On the privacy effect of data enhancement via the lens of memorization. *IEEE Transactions on Information Forensics and Security* 19 (2024), 4686–4699.
[45] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. 2022. Membership Inference Attacks by Exploiting Loss Trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2085–2098.
[46] Eugenio Lomurno and Matteo Matteucci. 2022. 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, 223–238.
[47] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Skq89Scxx
[48] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 346–363.
[49] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting Unintended Feature Leakage in Collaborative Learning. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 691–706.
[50] Sasi Kumar Murakonda and Reza Shokri. 2020. ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning. *arXiv e-prints* (2020), arXiv–2007.
[51] Vivek H Murthy, Harlan M Krumholz, and Cary P Gross. 2004. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *Jama* 291, 22 (2004), 2720–2726.
[52] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 739–753.
[53] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.
[54] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. 2018. Sok: Security and privacy in machine learning. In *2018 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 399–414.
[55] Joseph Pollock, Igor Shilov, Euodia Dodd, and Yves-Alexandre de Montjoye. 2025. Free {Record-Level} Privacy Risk Evaluation Through {Artifact-Based} Methods. In *34th USENIX Security Symposium (USENIX Security 25)*. 5525–5544.
[56] Zitao Qiao, Jinyuan Li, Michael Backes, and Yang Zhang. 2024. 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)*.
[57] Arezoo Rajabi, Reeya Pimple, Aiswarya Janardhanan, Surudhi Asokraj, Bhaskar Ramasubramanian, and Radha Poovendran. 2024. 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*. 1937–1939.
[58] Shahbaz Rezaei and Xin Liu. 2021. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7892–7900.
[59] Maria Rigaki and Sebastian Garcia. 2023. A survey of privacy attacks in machine learning. *Comput. Surveys* 56, 4 (2023), 1–34.
[60] Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. 2019. White-box vs Black-box: Bayes Optimal Strategies for Membership Inference. In *ICML 2019-36th International Conference on Machine Learning*, Vol. 97. 5558–5567.
[61] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. 2019. 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.
[62] Haonan Shi, Tu Ouyang, and An Wang. 2024. Learning-Based Difficulty Calibration for Enhanced Membership Inference Attacks. In *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 62–77.
[63] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.
[64] Leslie N Smith and Nicholay Topin. 2019. 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, 369–386.
[65] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* 25 (2012).
[66] Liwei Song and Prateek Mittal. 2021. Systematic Evaluation of Privacy Risks of Machine Learning Models. In *30th USENIX security symposium (USENIX security 21)*. 2615–2632.
[67] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.
[68] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. 2011. The German traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*. IEEE, 1453–1460.
[69] Elham Tabassi, Kevin J Burns, Michael Hadjimichael, Andres D Molina-Markham, and Julian T Sexton. 2019. A taxonomy and terminology of adversarial machine learning. *NIST IR* 2019 (2019), 1–29.
[70] Mingxing Tan and Quoc Le. 2021. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*. PMLR, 10096–10106.
[71] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*. 601–618.
[72] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2021. Demystifying Membership Inference Attacks in Machine Learning as a Service. *IEEE Transactions on Services Computing* 14, 06 (2021), 2073–2089.
[73] L. Del Vasto-Terrientes, D. Sánchez, and J. Domingo-Ferrer. 2025. Differential privacy in practice: lessons learned from 10 years of real-world applications. *IEEE Security & Privacy* In press (2025).
[74] Binghui Wang and Neil Zhenqiang Gong. 2018. Stealing hyperparameters in machine learning. In *2018 IEEE symposium on security and privacy (SP)*. IEEE, 36–52.
[75] Yanling Wang, Qian Wang, Lingchen Zhao, and Cong Wang. 2023. Differential privacy in deep learning: Privacy and beyond. *Future Generation Computer Systems* 148 (2023), 408–424.
[76] Zhiqi Wang, Chengyu Zhang, Yuetian Chen, Nathalie Baracaldo, Swanand Kadhe, and Lei Yu. 2025. Membership Inference Attacks as Privacy Tools: Reliability, Disparity and Ensemble. *arXiv preprint arXiv:2506.13972* (2025). To appear at ACM CCS 2025.
[77] Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. 2022. On the Importance of Difficulty Calibration in Membership Inference Attacks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=3eIrli0TwQ
[78] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. Enhanced Membership Inference Attacks against Machine Learning Models. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*. 3093–3106.
[79] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE Computer Society, 268–282.
[80] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. 2019. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5704–5713.
[81] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6023–6032.
[82] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide Residual Networks. In *British Machine Vision Conference 2016*. British Machine Vision Association.
[83] Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. 2024. Low-Cost High-Power Membership Inference Attacks. In *International Conference on Machine Learning*. PMLR, 58244–58282.
[84] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1Ddp1-Rb
[85] Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. 2025. 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, 333–345. https://doi.org/10.1109/SaTML64287.2025.00025
[86] Qiankun Zhang, Di Yuan, Boyu Zhang, Bin Yuan, and Bingqian Du. 2024. 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*. 1256–1270.
[87] Meiyi Zhu, Caili Guo, Chunyan Feng, and Osvaldo Simeone. 2025. On the Impact of Uncertainty and Calibration on Likelihood-Ratio Membership Inference Attacks. *IEEE Transactions on Information Forensics and Security* (2025).

A Additional Results on Reproducibility

This section reports additional reproducibility results.

Reproducibility without zero-FP filtering. We focus on *all* online LiRA detections on CIFAR-10 (AOF) (*i.e.*, without zero-FP filtering) for both target FPRs. The patterns align with the results reported

in the main paper: reproducibility drops rapidly as more runs are combined, and medium support thresholds improve stabilization.

Fig. 9 shows results at a target FPR of 0.001%, where agreement fell rapidly and coverage shrank with increasing support thresholds.

At the more tolerant 0.1% FPR (Fig. 10), stability increased slightly, but the union remained high for higher Jaccard similarities and continued to increase as more runs were combined.

Fig. 11 compares the top-16 most vulnerable samples across three seeds at two FPR levels (0.001% vs. 0.1%). There is clear cross-FPR inconsistency: in at least half of the runs, the identity and ranking of the top samples differ between the two FPRs. This confirms that even "top vulnerable samples" are not stable across FPR choices and that vulnerability assessments depend strongly on random training factors and threshold calibration.

Fig. 12 further shows that at 0.1% FPR, the online LiRA variant no longer produced zero-FP detections once the number of combined runs exceeded 10.

Table 13 reports reproducibility metrics for the alternative `mean gap` vulnerability score. The qualitative trends mirror those observed for the `median gap` in the main text, but stability in the extreme tail is consistently weaker.

At $q = 1\%$, pairwise tail Spearman drops from 53.45% (median gap) to 45.26% (mean gap), and variance increases (std. 20.49 vs. 13.79). The results are weaker, but qualitatively consistent with the `median gap`. Under strict multi-run agreement ($k = 12$), the globally stable core is also smaller. For example, at $q = 2.24\%$, the mean gap yields a global intersection of 338 samples (union 2466, Jaccard 13.71%), compared to 368 samples (union 2280, Jaccard 16.14%) under the median gap.

These differences indicate that the mean aggregation is more sensitive to variability across shadow models, leading to greater dispersion in the extreme tail. In contrast, the median gap is more robust to outlier shadow likelihood ratios, resulting in a more stable ranking signal across runs. Overall, while both scores preserve a population-level ordering, the median gap provides systematically stronger extreme-tail stability and tighter global agreement.

To understand the source of instability at the score level, Figure 13 plots the `median gap` score distribution within rank-percentile bins. Score dispersion was highest in the top 1–2% bins and decreased rapidly beyond 5–10%. In this sparse and high-variance regime, small training perturbations can reshuffle the ordering, whereas denser regions exhibit more stable rankings.

Fig. 14 selects, for each run, the top-1 most vulnerable sample (`median gap`) and evaluates its score across all 12 runs. We observe substantial cross-run dispersion, with large interquartile ranges and strong right-skew; in several cases, the median is far smaller than the mean, indicating that extreme values occur only in a subset of runs. Thus, even the most vulnerable record in a given run does not consistently exhibit extreme vulnerability magnitude under retraining variability.

Fig. 15 shows the top-1 sample from each of the 12 runs (`median gap`), arranged in a 3×4 grid.

B 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 14 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 mitigated 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 lowered 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 (5×10^{-4} to 10^{-2}). Lower values (5×10^{-5}) provided insufficient regularization (loss ratio 17.6), while excessive decay (5×10^{-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.

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 15, TL reduced LiRA's TPR by more than an order of magnitude while improving loss ratio and accuracy. We also observed that pretrained EfficientNet-V2 yields a comparable drop in attack success, but achieves higher test accuracy. Consequently, we employed EfficientNet-V2 for the main experiments, as a pragmatic practitioner will prefer higher model utility.

C Additional Training Details

Baseline vision models followed LiRA [11] with data augmentations consisting of random horizontal flip, reflection-padded 4px crop, and Cutout [17]; 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 hue jitter of ±0.1 (parameters reduced by half for GTSRB due to its more constrained color distribution); Cutout [17]

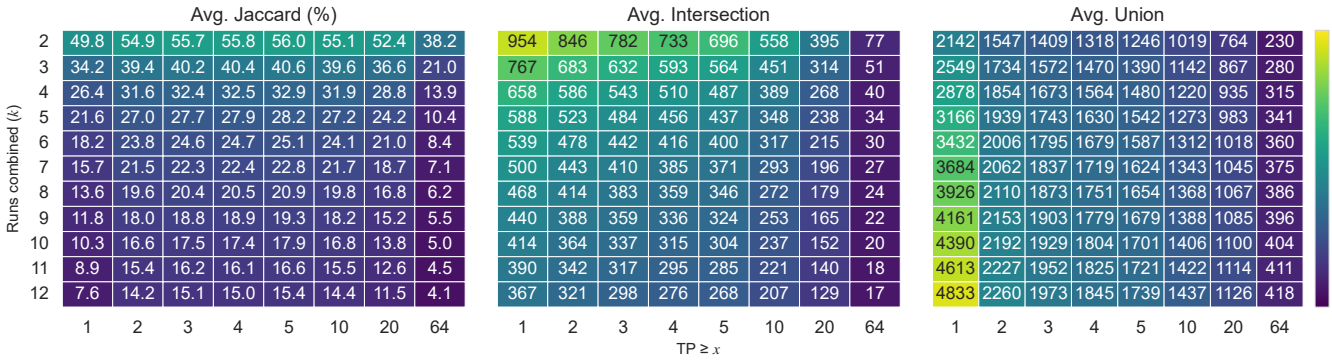


Figure 9: Reproducibility, stability, and coverage of *all* LiRA positives across runs (identical settings) at a target FPR of 0.001%.

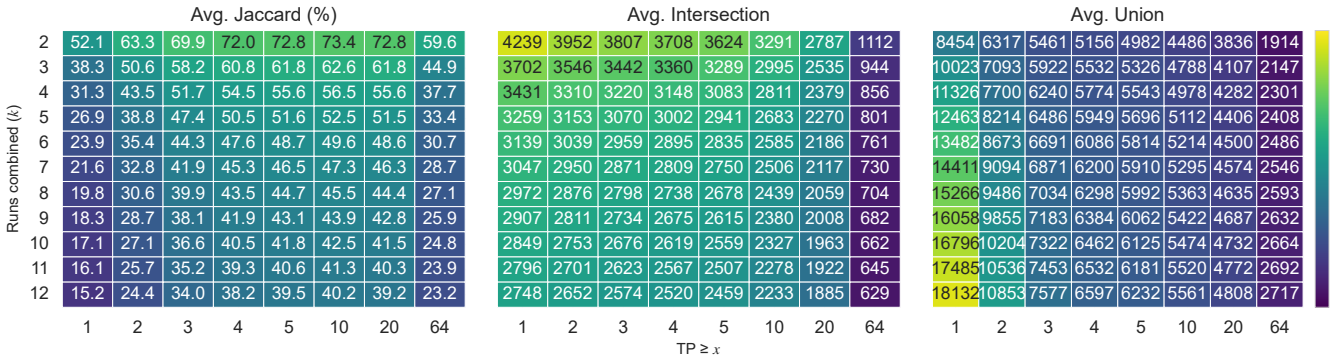


Figure 10: Reproducibility, stability, and coverage of *all* LiRA positives across runs (identical settings) at a target FPR of 0.1%.

(16×16 patches) for baseline configurations; MixUp [84] that linearly interpolates pairs of training samples and their labels with mixing coefficient $\lambda \sim \text{Beta}(1.0, 1.0)$; CutMix [81] that replaces rectangular regions between image pairs while mixing labels proportionally to area, applied with 50% probability and $\lambda \sim \text{Beta}(1.0, 1.0)$. 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 Appendix B). 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).

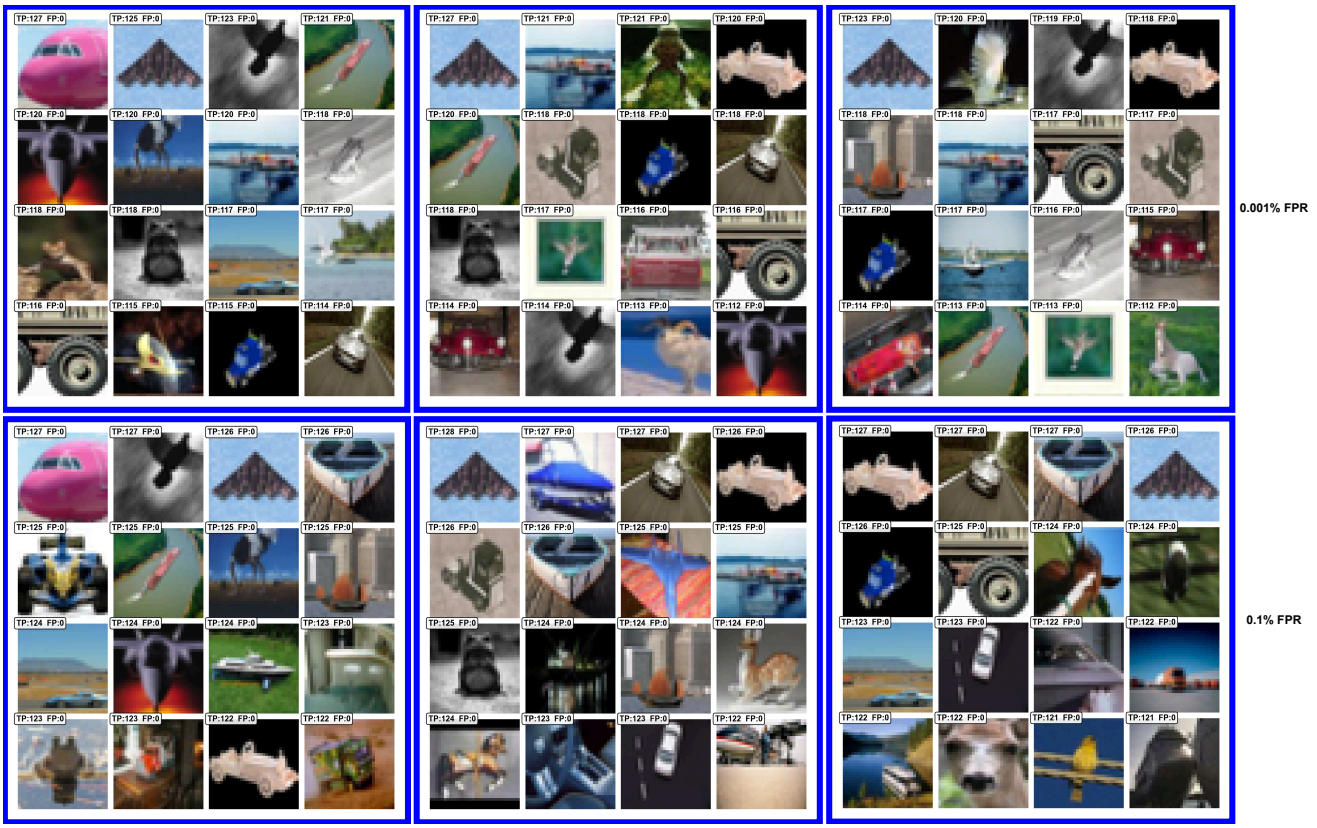


Figure 11: Top-16 most vulnerable samples across three seeds (columns) under two FPR settings (rows). Substantial cross-FPR disagreement is visible: many samples appear in only one FPR setting, and those shared often change rank.

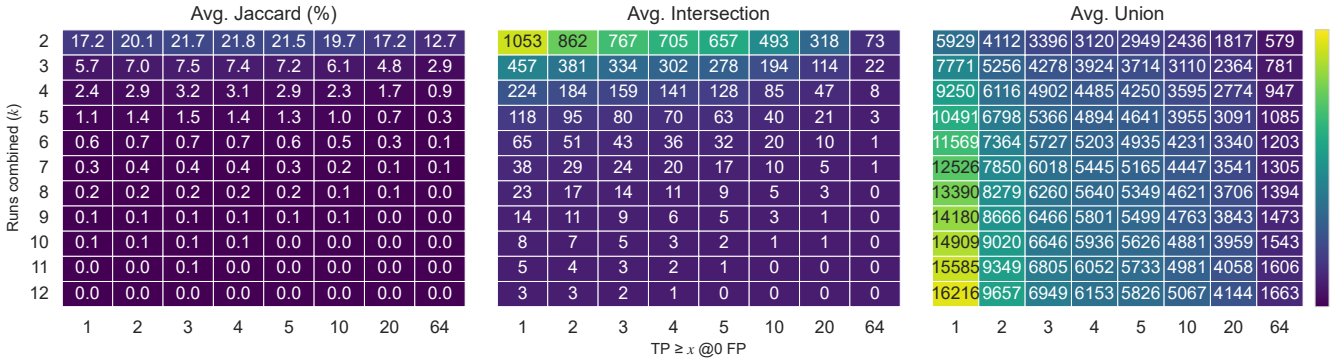


Figure 12: Reproducibility, stability, and coverage of zero-FP LiRA positives across runs (identical settings) for 0.1%FPR.

Table 13: Reproducibility of ranking-based sets (Top-q% by `mean gap` and FPR-thresholded sets)

		Pairwise ($k = 2$ runs)				All-runs ($k = 12$ runs)			
Setting	T_q	Tail ρ (%)	$\lvert \cap \rvert$	$\lvert \cup \rvert$	Jaccard (%)	Tail ρ (%)	$\lvert \cap \rvert$	$\lvert \cup \rvert$	Jaccard (%)
0.10	50	$39.41_{\pm 26.57}$	$23_{\pm 9}$	$77_{\pm 9}$	$31.57_{\pm 16.19}$	55.45	4	171	2.34
0.25	125	$43.13_{\pm 17.97}$	$65_{\pm 21}$	$185_{\pm 21}$	$37.31_{\pm 16.84}$	52.66	16	364	4.40
0.50	250	$45.16_{\pm 19.14}$	$137_{\pm 36}$	$363_{\pm 36}$	$39.20_{\pm 15.00}$	47.11	42	708	5.93
0.75	375	$44.03_{\pm 20.75}$	$222_{\pm 48}$	$528_{\pm 48}$	$43.17_{\pm 13.82}$	46.10	74	975	7.59
1.00	500	$45.26_{\pm 20.49}$	$305_{\pm 61}$	$695_{\pm 61}$	$45.09_{\pm 13.49}$	46.37	111	1255	8.84
2.00	1000	$50.68_{\pm 17.86}$	$677_{\pm 103}$	$1323_{\pm 103}$	$52.12_{\pm 12.38}$	50.35	296	2228	13.29
2.24	1121	$52.01_{\pm 17.50}$	$768_{\pm 110}$	$1474_{\pm 110}$	$53.03_{\pm 11.86}$	50.63	338	2466	13.71
3.00	1500	$54.92_{\pm 16.88}$	$1056_{\pm 137}$	$1944_{\pm 137}$	$55.06_{\pm 11.17}$	53.99	483	3153	15.32
4.00	2000	$57.92_{\pm 15.66}$	$1464_{\pm 170}$	$2536_{\pm 170}$	$58.42_{\pm 10.81}$	57.64	717	3931	18.24
5.00	2500	$59.80_{\pm 14.97}$	$1882_{\pm 201}$	$3118_{\pm 201}$	$61.02_{\pm 10.47}$	58.42	987	4708	20.96
10.00	5000	$69.03_{\pm 12.33}$	$4048_{\pm 332}$	$5952_{\pm 332}$	$68.51_{\pm 9.19}$	67.57	2459	8300	29.63
10.38	5190	$69.59_{\pm 12.17}$	$4211_{\pm 345}$	$6169_{\pm 345}$	$68.76_{\pm 9.21}$	68.04	2551	8572	29.76
20.00	10000	$77.69_{\pm 10.46}$	$8624_{\pm 532}$	$11376_{\pm 532}$	$76.18_{\pm 7.99}$	76.75	5902	14671	40.23
30.00	15000	$82.75_{\pm 9.38}$	$13141_{\pm 739}$	$16859_{\pm 739}$	$78.28_{\pm 7.59}$	81.62	9409	21296	44.18
40.00	20000	$85.86_{\pm 7.89}$	$17537_{\pm 1076}$	$22463_{\pm 1076}$	$78.45_{\pm 8.11}$	85.15	12588	28556	44.08
50.00	25000	$87.70_{\pm 6.61}$	$21537_{\pm 1431}$	$28463_{\pm 1431}$	$76.07_{\pm 8.25}$	87.29	15211	38180	39.84
60.00	30000	$88.95_{\pm 5.47}$	$25074_{\pm 1544}$	$34926_{\pm 1544}$	$72.10_{\pm 7.15}$	88.58	17142	46914	36.54
70.00	35000	$89.67_{\pm 4.59}$	$28987_{\pm 1397}$	$41013_{\pm 1397}$	$70.86_{\pm 5.56}$	89.31	18717	49776	37.60
80.00	40000	$89.10_{\pm 4.47}$	$34151_{\pm 943}$	$45849_{\pm 943}$	$74.56_{\pm 3.51}$	89.92	20882	49997	41.77
90.00	45000	$86.28_{\pm 6.18}$	$41099_{\pm 386}$	$48901_{\pm 386}$	$84.06_{\pm 1.46}$	90.15	25911	50000	51.82
100.00	50000	$80.46_{\pm 10.45}$	$50000_{\pm 0}$	$50000_{\pm 0}$	$100.00_{\pm 0.00}$	80.46	50000	50000	100.00
FPR-thresholded sets									
FPR = 0.001%	—	—	954	2142	49.80	—	367	4833	7.60
FPR = 0.1%	—	—	4239	8454	52.10	—	2748	18132	15.20

Table 14: 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 15: 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

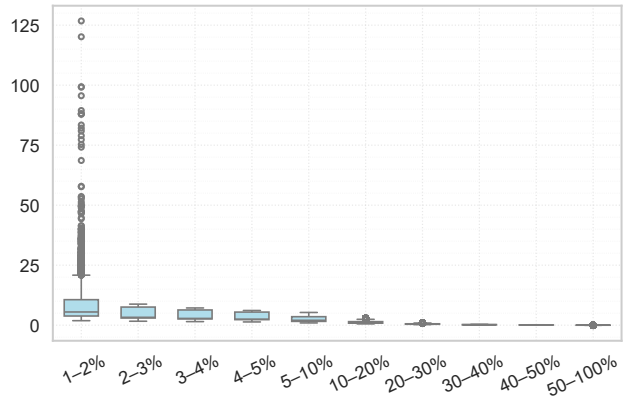


Figure 13: Distribution of `median gap` vulnerability scores within rank-percentile bins (bins computed per run; values pooled over 12 independent runs).

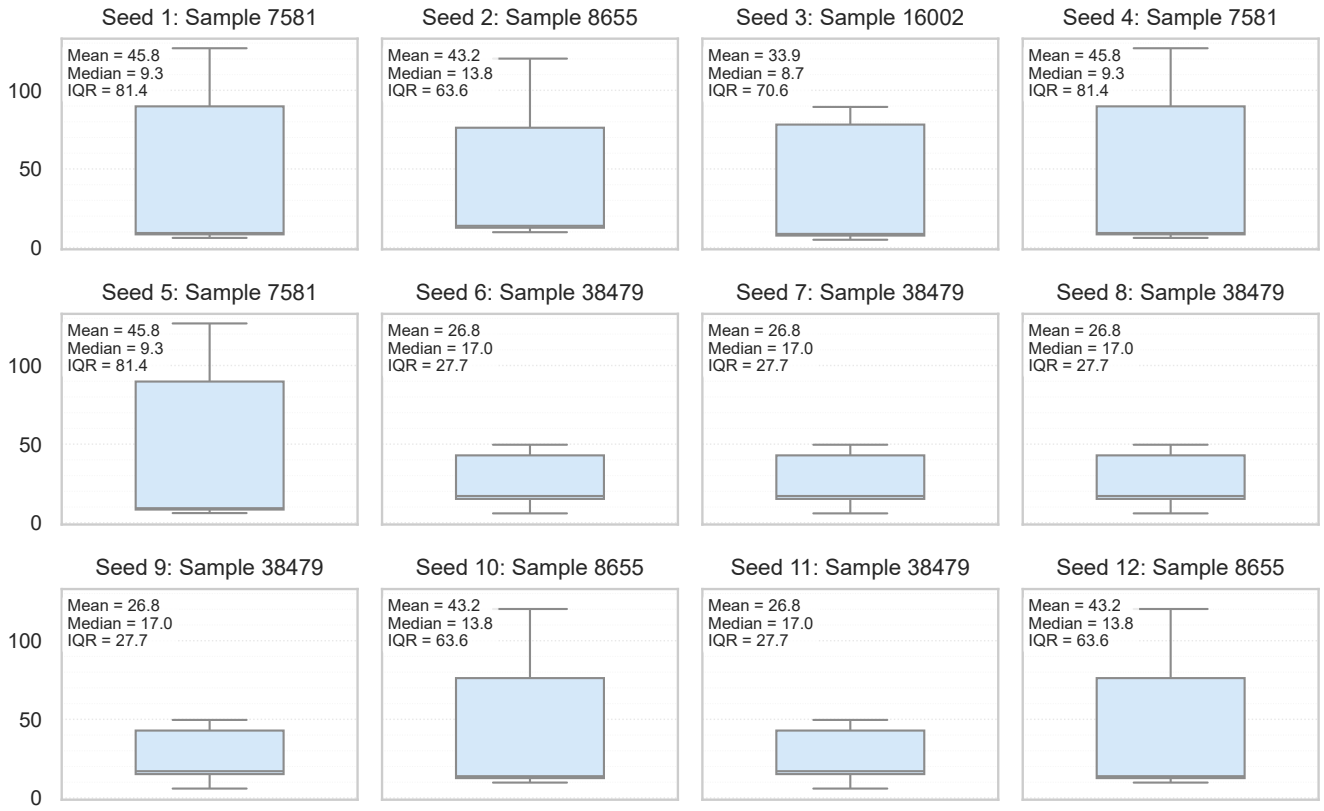


Figure 14: For each run, the top-1 most vulnerable sample is selected, and its `median gap` score is evaluated across all 12 retrainings. Large interquartile ranges and strong right-skew indicate substantial cross-run variance in extreme vulnerability magnitude. While the top-1 identity rotates among a small set of recurring candidates, their scores vary markedly across runs, confirming sensitivity at the extreme boundary.

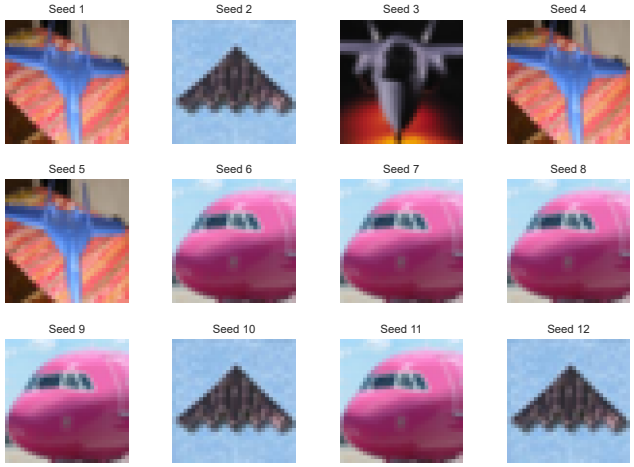


Figure 15: Top-1 identity across runs. For each seed (panel), we display the single most vulnerable sample in that run (highest `median gap`). Repeated appearances of the same image across seeds indicate overlap in the extreme tail.