MA-LiRA: MA-LiRA: Memory-Aware Membership Inference Attacks

Anonymous authors
Paper under double-blind review

Abstract

Membership inference asks whether a candidate example participated in a model's training set and provides an important tool for empirical privacy auditing. Many strong attacks rely on statistics observed only at the final model checkpoint, potentially discarding information contained in how a candidate's behavior evolves during optimization. We introduce MA-LiRA, a checkpoint-aware membership-inference attack that exploits this training history. For each labeled candidate, MA-LiRA constructs a compact five-dimensional representation from its checkpoint-wise loss trajectory, capturing endpoint fit, overall learning progress, local trajectory irregularity, and temporally weighted changes. Each feature is calibrated using candidate-specific member and non-member shadow distributions, and the resulting likelihood-ratio coordinates are combined by a fusion model trained exclusively from shadow-derived pseudo-target examples, without using target membership labels for fitting or threshold selection. We evaluate MA-LiRA on CIFAR-10, CIFAR-100, and SVHN against a broad set of score-based, calibrated, likelihood-ratio, and trajectory-based membership-inference baselines. Across the evaluated datasets, MA-LiRA consistently improves overall membership discrimination and low-false-positive-rate detection under the stated checkpoint-aware access model. These results indicate that training-history information can provide membership evidence beyond that available from a final-checkpoint statistic alone.

1 Introduction

A trained model can reveal whether a particular record was used to fit it. Membership-inference attacks formalize this question and expose risks ranging from disclosure of participation in a sensitive study to memorization of unusual examples (Shokri et al., 2017; Yeom et al., 2018). The same attacks also serve a constructive role: a model owner can use them to audit empirical privacy exposure before deployment. This role demands more than aggregate attack accuracy. When non-members greatly outnumber members, even a seemingly accurate attack can generate many false alarms; conversely, a small set of high-confidence true positives can represent material leakage while barely affecting average accuracy (Carlini et al., 2022). We therefore treat low-false-positive-rate performance as a primary object of study.

Many attacks obtain evidence from the final model's loss, confidence, entropy, difficulty-adjusted score, or likelihood ratio (Yeom et al., 2018; Sablayrolles et al., 2019; Watson et al., 2022; Carlini et al., 2022). These statistics are effective because training often changes a model's behavior on members relative to comparable non-members. Yet a final checkpoint is a compressed view of optimization: two candidates can end at similar losses after following different paths. Prior work does not uniformly ignore this history. In particular, TrajectoryMIA constructs a distilled loss trajectory from a final teacher model and trains an attack classifier on that surrogate path together with the teacher loss (Liu et al., 2022). Its result motivates a sharper access-dependent question rather than a claim that trajectories have not been studied.

We consider a checkpoint-aware audit in which the auditor can evaluate labeled candidates at actual training checkpoints. This setting is natural for model developers, internal red teams, and regulated audit pipelines that retain training artifacts, but it is stronger than final-model-only black-box access. It is also distinct from distillation-based trajectory reconstruction: the observations here track the target optimization itself. Our central question is: *when training checkpoints are available for privacy*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

auditing, can checkpoint-wise loss evolution improve per-example likelihood-ratio inference beyond a final-checkpoint statistic?

MA-LiRA addresses this question in three stages. It evaluates each candidate at retained training checkpoints and summarizes its loss history using five complementary endpoint, trajectory, and memory-aware statistics. Each feature is calibrated separately using candidate-specific member and non-member shadow distributions, and the resulting log-likelihood ratios are standardized and combined using fusion learned exclusively from shadow data.

Our contributions are threefold. First, we introduce a compact checkpoint-aware representation that combines endpoint, trajectory, and memory-aware information from per-example loss histories. Second, we develop feature-wise LiRA calibration with shadow-only learned fusion, avoiding target-membership-label tuning. Third, we provide a comprehensive evaluation on CIFAR-10, CIFAR-100, and SVHN against classical score-based, shadow-model, calibrated reference, likelihood-ratio, quantile-regression, and trajectory-based attacks, complemented by endpoint-statistic, class-wise, and difficulty-conditioned analyses with explicit attention to low-FPR resolution and access assumptions.

2 Related Work

Classical membership inference. Shokri et al. (2017) trained shadow models and class-conditional attack models on prediction vectors, establishing a practical black-box attack. Yeom et al. (2018) related leakage to overfitting and influence and proposed a loss-threshold rule. Subsequent work has considered loss, true-label confidence, maximum confidence, entropy, margins, and learned combinations (Song & Mittal, 2021). Population-level overfitting can create membership signal, but example-specific difficulty means that candidates with similar endpoint scores need not provide equivalent membership evidence (Yeom et al., 2018; Watson et al., 2022; Carlini et al., 2022).

Calibration and likelihood ratios. Sablayrolles et al. (2019) derived Bayes-optimal membership rules under modeling assumptions and motivated loss-based approximations, while Watson et al. (2022) showed that calibrating against example difficulty can materially improve practical attacks. LiRA fits per-example IN and OUT score distributions and evaluates a likelihood ratio, with particular attention to very low FPRs (Carlini et al., 2022). RMIA constructs a population-relative test designed to retain power with fewer reference models (Zarifzadeh et al., 2024), and InfoRMIA replaces RMIA's discrete population comparison with an information-theoretic continuous statistic (Tao & Shokri, 2026). Quantile-MIA amortizes calibration through a conditional quantile model trained on auxiliary non-members (Bertran et al., 2023). Together, these methods motivate candidate-specific calibration rather than a single global threshold.

Trajectory-based inference. TrajectoryMIA reconstructs a surrogate loss trajectory by distilling final target and shadow teachers and trains an attack classifier on the resulting trajectory together with the teacher loss (Liu et al., 2022). It therefore requires only final-model queries, but incurs auxiliary-data and distillation costs. In contrast, MA-LiRA directly observes actual target and shadow training checkpoints and combines structured temporal summaries with feature-wise likelihood-ratio calibration and shadow-only fusion. The two methods consequently operate under different access assumptions and should be interpreted accordingly.

3 Methodology

3.1 Problem Setting and Threat Model

Let $\mathcal{X}$ and $\mathcal{Y} = \{1, \dots, C\}$ denote the input and label spaces. A private training set $D_{\text{tr}} \subset \mathcal{X} \times \mathcal{Y}$ is used to optimize a classifier f_θ. The retained parameter sequence is $\Theta = (\theta_1, \dots, \theta_T)$, where θ_t is the checkpoint at the tth retained training stage and f_{θ_T} is the final target model. For a labeled candidate $z = (x, y)$, define $m(z) = 1$ if $z \in D_{\text{tr}}$ and $m(z) = 0$ otherwise. An attack produces a continuous score $S(z)$, with larger values indicating stronger membership evidence, ideally approximating

$$S(z) \approx \log \frac{p(\mathcal{O}(z) \mid m(z) = 1)}{p(\mathcal{O}(z) \mid m(z) = 0)}, \tag{1}$$

where $\mathcal{O}(z)$ contains the checkpoint-wise observations available to the auditor.

The auditor knows y, can evaluate z at selected target checkpoints, and can train B shadow models on samples from a related population. Shadow s has its own aligned checkpoints $\Theta^{(s)}$ and a known audit membership mask, which supplies attack-training labels. The implementation loads checkpoint files locally and evaluates logits; it uses raw, unaugmented-view cross-entropy losses at every checkpoint. Equivalent query access to the true-label probability at each checkpoint would suffice, but access only to the final target model would not. The target membership labels are unavailable during attack construction and may not be used to orient scores, choose hyperparameters or attack epochs, select thresholds, or fit any component. They are revealed only after scores and operating rules are frozen, for metric computation.

The audit population contains a balanced set of target members and disjoint target non-members in the current implementation. Each shadow independently includes an audit candidate with probability approximately one half and fills the remainder of its fixed-size training set from an eligible pool. This produces candidate-specific IN and OUT reference observations. Separate auxiliary partitions are reserved for RMIA, Quantile-MIA, and TrajectoryMIA; Section 4 specifies the contamination controls. Shadow count B, checkpoint count T, and memory length $T-1$ are distinct quantities.

3.2 MA-LiRA Overview and Checkpoint-Wise Loss

The MA-LiRA pipeline is

checkpoints $\rightarrow$ loss trajectory $\rightarrow$ feature bank $\rightarrow$ feature-wise likelihood ratios $\rightarrow$ fusion $\rightarrow S(z)$.

During attack training, shadow checkpoints provide labeled pseudo-target observations from which MA-LiRA learns feature calibrators and a fusion rule. During target auditing, these frozen components transform target checkpoint losses into a membership score.

Let $p_t(x) = \mathrm{softmax}(f_{\theta_t}(x))$. The implementation uses the unmodified per-example cross-entropy

$$\ell_t(z) = -\log[p_t(x)]_y, \qquad \boldsymbol{\ell}(z) = (\ell_1(z), \ldots, \ell_T(z)), \tag{2}$$

with increments

$$\Delta\ell_t(z) = \ell_{t+1}(z) - \ell_t(z), \qquad t = 1, \ldots, T-1.$$

Negative increments indicate decreasing candidate loss and positive increments indicate increasing loss. The first retained point is the first saved training checkpoint, not an untrained initialization. Checkpoints are sorted by their encoded epoch, and target and shadows must have identical epoch lists.

3.3 The Five-Dimensional Feature Bank

For each example $z = (x, y)$, MA-LiRA constructs a five-dimensional feature vector from its retained loss trajectory $\{\ell_t(z)\}_{t=1}^{T}$. The first two coordinates are the final loss $\ell_T(z)$ and the net loss drop $\ell_1(z) - \ell_T(z)$. For $T \geq 3$, trajectory roughness is

$$R(z) = \frac{1}{T-2} \sum_{t=1}^{T-2} \left|\Delta\ell_{t+1}(z) - \Delta\ell_t(z)\right|, \tag{3}$$

which measures local irregularity in the loss path. The remaining two coordinates are the signed and absolute power-law memory summaries defined in Section 3.4. The complete ordered feature vector is

$$\mathbf{F}_\alpha(z) = \left[\ell_T(z),\ \ell_1(z) - \ell_T(z),\ R(z),\ M_\alpha^{\mathrm{sgn}}(z),\ M_\alpha^{\mathrm{abs}}(z)\right]^\top \in \mathbb{R}^5. \tag{4}$$

Thus the compact bank contains one endpoint statistic, two unweighted trajectory summaries, and two explicitly memory-aware summaries, capturing final fit, global improvement, local oscillation, temporal direction, and temporal variation without the redundancy of a larger handcrafted feature bank.

3.4 POWER-LAW TEMPORAL MEMORY

For increment t, define its index age relative to the most recent increment as

$$r_t = (T-1) - t.$$

With $0 < \alpha < 1$, the implemented power-law weight is

$$w_t^{(\alpha)} = (r_t + 1)^{\alpha - 1}, \qquad t = 1, \ldots, T-1. \tag{5}$$

The two memory-aware coordinates are

$$M_\alpha^{\mathrm{sgn}}(z) = \frac{\sum_{t=1}^{T-1} w_t^{(\alpha)} \Delta\ell_t(z)}{\sum_{t=1}^{T-1} w_t^{(\alpha)}}, \tag{6}$$

$$M_\alpha^{\mathrm{abs}}(z) = \frac{\sum_{t=1}^{T-1} w_t^{(\alpha)} |\Delta\ell_t(z)|}{\sum_{t=1}^{T-1} w_t^{(\alpha)}}. \tag{7}$$

The signed statistic captures the recency-weighted direction of loss change, whereas the absolute statistic captures recency-weighted variation irrespective of direction. Recent transitions have smaller age and therefore larger weight, while older transitions remain present. Normalization reduces dependence on checkpoint count. The current configuration fixes $\alpha = 0.7$; it is not learned by the fusion classifier. As $\alpha \to 1$, the weights approach a uniform kernel; as $\alpha \to 0$, they approach reciprocal-age decay.

When $T < 2$, the implementation sets the memory features to zero; when $T < 3$, it also sets roughness to zero. Constant paths yield zero change-based features. Cross-entropy values are evaluated in finite precision, and downstream variance floors handle collapsed reference distributions. Missing or repeated checkpoint epochs are not imputed: repeated common epochs are rejected, inconsistent target/shadow schedules are rejected, and the retained order is treated as equally spaced even if physical training intervals differ. Uniform and exponential kernels are controlled sensitivity variants, not separate proposed attacks.

3.5 FEATURE-WISE LIKELIHOOD-RATIO CALIBRATION

Let ϕ_{sij} be feature j for candidate i under shadow model s, and let m_{si} indicate whether candidate i was in shadow s's training set. For each candidate and coordinate, MA-LIRA fits candidate-specific Gaussian models

$$q_{ij,c}(u) = \mathcal{N}(u; \mu_{ij,c}, \sigma^2_{ij,c}), \quad c \in \{\mathrm{in}, \mathrm{out}\}, \qquad \lambda_{ij} = \log \frac{q_{ij,\mathrm{in}}(\phi_{ij})}{q_{ij,\mathrm{out}}(\phi_{ij})}. \tag{8}$$

The means and population variances use shadows satisfying $m_{si} = 1$ or $m_{si} = 0$, respectively. Computed variances are clipped below at zero and then increased by $\varepsilon = 10^{-8}$; the Gaussian density again applies the same lower bound. If fewer than two observations exist on either side, that side falls back to a global Gaussian pooled over all candidate–shadow observations with the corresponding membership state. If either global side has fewer than two observations, both fall back to the flattened reference bank. No additional likelihood-ratio clipping is applied. The orientation in Equation 8 makes larger coordinates favor membership.

To generate fusion-training data without self-calibration, each shadow is held out as a pseudo-target while the remaining $B-1$ shadows estimate its candidate-specific IN and OUT distributions. Stacking all held-out shadows yields

$$X_{\mathrm{LOO}} \in \mathbb{R}^{BN \times 5}$$

and known shadow membership labels $\boldsymbol{m}_{\mathrm{LOO}}$. The real target never participates in this construction. Stacking all held-out shadows yields $\mathbf{m}_{\mathrm{LOO}} \in \{0,1\}^{BN}$, where $\mathbf{m}_{\mathrm{LOO}}$ contains the known membership labels of the leave-one-shadow-out pseudo-target examples. The real target never participates in this construction.

3.6 SHADOW-ONLY LEARNED FUSION AND THRESHOLDING

Let $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{i,5})$. A standardizer fitted on X_{LOO} forms

$$\widetilde{\boldsymbol{\lambda}}_i = (\boldsymbol{\lambda}_i - \boldsymbol{\mu}_\lambda) \oslash \boldsymbol{s}_\lambda,$$

Algorithm 1: Memory-Aware Likelihood-Ratio Attack (MA-LiRA)

Input: shadow checkpoints $\{\Theta^{(s)}\}_{s=1}^{B}$, shadow masks $\{m_{si}\}$, target checkpoints Θ, and labeled candidates $\{z_i\}_{i=1}^{N}$.

Shadow phase.

1. Compute shadow loss trajectories and $\mathbf{F}_{\alpha}^{(s)}(z_i) \in \mathbb{R}^5$ using Eqs. (2)–(7).
2. For each held-out shadow s, estimate candidate-wise IN/OUT Gaussians from the remaining $B-1$ shadows and compute $\boldsymbol{\lambda}_{si}$ using Eq. (8).
3. Stack the LOSO likelihood-ratio vectors, standardize them, and fit the shadow-only fusion in Eq. (9).
4. Select $\tau^\star$ on a held-out calibration pseudo-target using Eq. (10), when required.

Target phase.

1. Compute the target loss trajectory and $\mathbf{F}_{\alpha}(z_i)$ using Eqs. (2)–(7).
2. Using all shadows, estimate the five candidate-wise IN/OUT Gaussian pairs and compute $\boldsymbol{\lambda}_i$ with Eq. (8).
3. Apply the frozen standardizer and fusion in Eq. (9) to obtain
$$S(z_i) = b + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{\lambda}}_i.$$
4. Return $S(z_i)$ and, when required,
$$\widehat{m}_i = \mathbb{I}S(z_i) \geq \tau^\star.$$

using unit scale for a constant coordinate. The default fusion is a class-balanced, ℓ_2-regularized logistic regression (L-BFGS, maximum 2000 iterations, default inverse regularization $C = 1$):

$$P_{\boldsymbol{\beta},b}(m = 1 \mid z) = \sigma\Big(b + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{\lambda}}(z)\Big), \qquad \sigma(a) = (1 + e^{-a})^{-1}. \tag{9}$$

The implementation reports the decision function $S(z) = b + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{\lambda}}(z)$, which is monotone in the predicted probability. The coefficients $\boldsymbol{\beta} \in \mathbb{R}^5$ and b are learned from shadow-derived examples. They are unrelated to α, checkpoint count, or memory length.

ROC metrics require no fixed threshold. For threshold-dependent metrics, a held-out shadow pseudo-target is used to select

$$\tau^\star = \arg\max_{\tau} \left[\mathrm{TPR}_{\mathrm{cal}}(\tau) - \mathrm{FPR}_{\mathrm{cal}}(\tau)\right]. \tag{10}$$

The calibration pseudo-target is excluded from fitting the temporary attack used to select $\tau^\star$. The real target is then scored by the final fusion trained on the full leave-one-shadow-out stack. Target membership labels are never used for threshold selection. Because the threshold transfers between two fitted fusion models, it is used only for secondary threshold-dependent metrics; primary ROC metrics require no threshold. Oracle thresholds computed on target labels are diagnostic only.

3.7 Complexity, Storage, and Implementation Controls

Let C_f be one forward-pass cost and P the number of model parameters. Collecting target and shadow observations costs $O((B+1)NTC_f)$ and storing all model checkpoints costs $O((B+1)TP)$; cached trajectories require $O((B+1)NT)$. Feature extraction is $O((B+1)NT)$. Target feature calibration is $O(BNd)$ for $d = 5$. The direct leave-one-shadow-out implementation recomputes calibrators and costs $O(B^2Nd)$ before fitting the fusion on BN rows; with I optimization passes, logistic fitting is approximately $O(IBNd)$. Once shadow artifacts are fitted, auditing N target candidates costs $O(NTC_f + BNd + Nd)$. This is more expensive than endpoint-only LiRA in both checkpoint evaluations and storage.

Derived caches carry a method version, dataset and class count, normalization and augmentation settings, seed, architecture, optimization parameters, split sizes, shadow budget, checkpoint schedule, and baseline-specific auxiliary settings. A cache whose metadata differs is rejected. Partition manifests store source-aware indices and hashes; model caches require aligned checkpoint epochs and deterministic loader ordering. These controls reduce accidental reuse across datasets or incompatible runs but do not replace independent replication.

Table 1: Cross-dataset comparison of membership inference attacks on CIFAR-10, CIFAR-100, and SVHN. Higher values are better for all metrics. Bold denotes the best result within each dataset, while underlining denotes the strongest baseline result.

Dataset	Reference / Method	AUC ↑	TPR@1% FPR ↑	TPR@0.1% FPR ↑	Shadow-Cal. Adv. ↑
CIFAR-10	Yeom et al. (2018)	0.6029	0.0112	0.0000	0.1714
	Song & Mittal (2021) (confidence)	0.5491	0.0112	0.0000	0.0724
	Song & Mittal (2021) (entropy)	0.5471	0.0112	0.0002	0.0772
	Shokri et al. (2017)	0.5953	0.0182	0.0020	0.1360
	Sablayrolles et al. (2019)	0.6576	0.0528	0.0210	0.2176
	Watson et al. (2022)	0.6443	0.0516	0.0190	0.1886
	Carlini et al. (2022)	0.6296	0.0348	0.0048	0.1734
	Zarifzadeh et al. (2024)	0.6393	0.0444	0.0078	0.1888
	Tao & Shokri (2026)	0.6393	0.0452	0.0086	0.1888
	Liu et al. (2022)	0.5730	0.0126	0.0038	0.1156
	MA-LiRA (Ours)	**0.7874**	**0.1584**	**0.0282**	**0.3940**
CIFAR-100	Yeom et al. (2018)	0.6052	0.0120	0.0000	0.1624
	Song & Mittal (2021) (confidence)	0.5530	0.0120	0.0000	0.0796
	Song & Mittal (2021) (entropy)	0.5517	0.0118	0.0002	0.0736
	Shokri et al. (2017)	0.5886	0.0174	0.0044	0.1412
	Sablayrolles et al. (2019)	0.6632	0.0558	0.0098	0.2224
	Watson et al. (2022)	0.6479	0.0548	0.0094	0.1948
	Carlini et al. (2022)	0.6379	0.0392	0.0056	0.1824
	Zarifzadeh et al. (2024)	0.6464	0.0548	0.0064	0.1924
	Tao & Shokri (2026)	0.6464	0.0548	0.0064	0.1904
	Liu et al. (2022)	0.5760	0.0144	0.0018	0.1246
	MA-LiRA (Ours)	**0.7984**	**0.1780**	**0.0472**	**0.4226**
SVHN	Yeom et al. (2018)	0.5152	0.0106	0.0012	0.0198
	Song & Mittal (2021) (confidence)	0.5123	0.0106	0.0012	0.0158
	Song & Mittal (2021) (entropy)	0.5123	0.0110	0.0012	0.0230
	Shokri et al. (2017)	0.5190	0.0096	0.0008	0.0304
	Sablayrolles et al. (2019)	0.5332	0.0246	0.0070	0.0384
	Watson et al. (2022)	0.5297	0.0246	0.0070	0.0362
	Carlini et al. (2022)	0.5345	0.0164	0.0018	0.0288
	Zarifzadeh et al. (2024)	0.5325	0.0218	0.0056	0.0344
	Tao & Shokri (2026)	0.5327	0.0230	0.0062	0.0352
	Liu et al. (2022)	0.5023	0.0086	0.0000	-0.0006
	MA-LiRA (Ours)	**0.5715**	**0.0474**	**0.0096**	**0.0770**

4 EXPERIMENTS

We focus on three questions: **RQ1**, does MA-LIRA improve overall and low-FPR membership inference across CIFAR-10, CIFAR-100, and SVHN? **RQ2**, are the gains consistent across CIFAR-100 classes rather than concentrated in a few categories? **RQ3**, do they persist under difficulty-conditioned member/non-member shifts? We evaluate CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) and SVHN (Netzer et al., 2011). Each main audit set contains 5,000 target members and 5,000 disjoint target non-members, and target membership labels are revealed only after attack scores and non-oracle operating rules are frozen. The confirmed CIFAR-100 configuration uses a small-image ResNet-18 (He et al., 2016), 30 training epochs, and $B = 32$ shadow models; CIFAR-10 and SVHN use the same audit protocol. We compare MA-LIRA with the loss attack of Yeom et al. (2018), confidence and entropy baselines from Song & Mittal (2021), and the attacks of Shokri et al. (2017), Sablayrolles et al. (2019), Watson et al. (2022), Carlini et al. (2022), Zarifzadeh et al. (2024), Tao & Shokri (2026), and Liu et al. (2022). The CIFAR-100 class-wise analysis additionally includes Bertran et al. (2023). External methods are identified by source authors in the tables, while internal MA-LIRA variants are labeled as ablations. The main MA-LIRA result is the raw-fused five-feature attack. Endpoint/reference baselines use the final target model, whereas MA-LIRA accesses retained target and shadow checkpoints. Liu et al. (2022) instead reconstructs a surrogate trajectory by distillation. All non-oracle thresholds are selected without target membership labels; target-label-selected thresholds are reported only as diagnostics. We report ROC-AUC, TPR at 1% and 0.1% FPR, and shadow-calibrated membership advantage $J = \mathrm{TPR} - \mathrm{FPR}$. ROC metrics are threshold-free, while shadow-calibrated advantage uses an operating point selected without target labels. Low-FPR results are interpreted only at empirically supported FPR resolutions; oracle accuracy-like metrics are diagnostic.

Cross-dataset performance. Across all three datasets, MA-LIRA achieves the highest value for every primary metric in Table 1. The advantage is strongest on CIFAR-10 and CIFAR-100, where AUC reaches 0.7874 and 0.7984 and TPR@1% FPR reaches 0.1584 and 0.1780; SVHN is more challenging overall, but MA-LIRA remains strongest in the reported comparison.

Class-wise consistency. At the class level, the CIFAR-100 improvement is broadly distributed rather than driven by a few categories (Table 2). MA-LIRA reaches 0.7891 mean AUC and ranks first in 94 of 100 classes, whereas the endpoint-only ablation drops to 0.5249.

Across all three difficulty-conditioned settings, MA-LIRA retains a substantial advantage in AUC and at 1% FPR (Table 3). The separation is especially pronounced for hard members, with AUCs of 0.8022 in H/A and 0.7107 in H/E. Results at 0.1% FPR are also reported for completeness, but in the

Table 2: Class-wise AUC summary on CIFAR-100. Statistics are computed over the 100 within-class AUCs. Class wins count the classes in which a method attains the highest AUC.

Reference / Method	Mean	Median	Min.	Max.	Wins
MA-LiRA (Ours)	**0.7891**	**0.7949**	**0.5594**	**0.9083**	**94/100**
Sablayrolles et al. (2019)	0.6669	0.6687	0.5044	0.8050	1/100
Watson et al. (2022)	0.6533	0.6543	0.4666	0.8013	2/100
Tao & Shokri (2026)	0.6501	0.6412	0.5473	0.7859	0/100
Zarifzadeh et al. (2024)	0.6500	0.6416	0.5473	0.7861	0/100
Carlini et al. (2022)	0.6330	0.6311	0.4459	0.7972	0/100
Yeom et al. (2018)	0.6131	0.6110	0.4852	0.7755	0/100
Bertran et al. (2023)	0.6126	0.6099	0.4699	0.7530	3/100
Shokri et al. (2017)	0.5954	0.5946	0.4314	0.7853	0/100
Liu et al. (2022)	0.5747	0.5713	0.4104	0.7308	0/100
Song & Mittal (2021) (entropy)	0.5480	0.5559	0.3855	0.6984	0/100
MA-LiRA (endpoint-only ablation)	0.5249	0.5344	0.3145	0.6486	0/100

Table 3: Difficulty-conditioned stress tests on CIFAR-100. Hard members are the 1,000 highest-loss target-training examples, whereas easy non-members are the 1,000 lowest-loss target-test examples. H/A compares hard members against all 5,000 non-members; A/E compares all 5,000 members against easy non-members; and H/E compares hard members against easy non-members. All metrics are higher-is-better. Oracle balanced accuracy and advantage use a target-audit threshold selected to maximize $\mathrm{TPR} - \mathrm{FPR}$ and are included only as diagnostic operating-point metrics.

Reference / Method	AUC	Oracle Bal. Acc.	Oracle Adv.	TPR@1% FPR	TPR@0.1% FPR
H/A: hard members vs. all non-members ($n_{\mathrm{M}} = 1{,}000, n_{\mathrm{N}} = 5{,}000$)					
Carlini et al. (2022)	0.4304	0.5104	0.0208	0.0160	0.0030
Zarifzadeh et al. (2024)	0.2332	0.5024	0.0048	0.0000	0.0000
Tao & Shokri (2026)	0.2332	0.5024	0.0048	0.0000	0.0000
Watson et al. (2022)	0.4398	0.5356	0.0712	0.0330	0.0090
Sablayrolles et al. (2019)	0.3383	0.5129	0.0258	0.0340	0.0060
Ours: MA-MIA shadow	0.4580	0.5356	0.0712	0.0190	0.0000
Liu et al. (2022)	0.3742	0.5128	0.0256	0.0000	0.0000
MA-LiRA (raw-sum ablation)	0.6883	0.6434	0.2868	0.0300	0.0030
MA-LiRA (Ours)	**0.8022**	**0.7322**	**0.4644**	**0.2720**	**0.0760**
A/E: all members vs. easy non-members ($n_{\mathrm{M}} = 5{,}000, n_{\mathrm{N}} = 1{,}000$)					
Carlini et al. (2022)	0.5074	0.5480	0.0960	0.0256	0.0044
Zarifzadeh et al. (2024)	0.3932	0.5506	0.1012	0.0604	0.0288
Tao & Shokri (2026)	0.3933	0.5509	0.1018	0.0604	0.0288
Watson et al. (2022)	0.6157	0.6443	0.2886	0.1476	0.0756
Sablayrolles et al. (2019)	0.5061	0.6023	0.2046	0.1340	0.0812
Ours: MA-MIA shadow	0.4451	0.5272	0.0544	0.0400	0.0154
Liu et al. (2022)	0.4595	0.5407	0.0814	0.0040	0.0000
MA-LiRA (raw-sum ablation)	0.6280	0.6062	0.2124	0.0068	0.0006
MA-LiRA (Ours)	**0.6945**	**0.6713**	**0.3426**	**0.2256**	**0.1038**
H/E: hard members vs. easy non-members ($n_{\mathrm{M}} = 1{,}000, n_{\mathrm{N}} = 1{,}000$)					
Carlini et al. (2022)	0.2900	0.5025	0.0050	0.0110	0.0030
Zarifzadeh et al. (2024)	0.0000	0.5000	0.0000	0.0000	0.0000
Tao & Shokri (2026)	0.0000	0.5000	0.0000	0.0000	0.0000
Watson et al. (2022)	0.3626	0.5790	0.1580	0.0880	0.0460
Sablayrolles et al. (2019)	0.1432	0.5265	0.0530	0.0580	0.0410
Ours: MA-MIA shadow	0.2064	0.5050	0.0100	0.0200	0.0050
Liu et al. (2022)	0.2987	0.5340	0.0680	0.0000	0.0000
MA-LiRA (raw-sum ablation)	0.5628	0.5965	0.1930	0.0120	0.0020
MA-LiRA (Ours)	**0.7107**	**0.7255**	**0.4510**	**0.3280**	**0.1720**

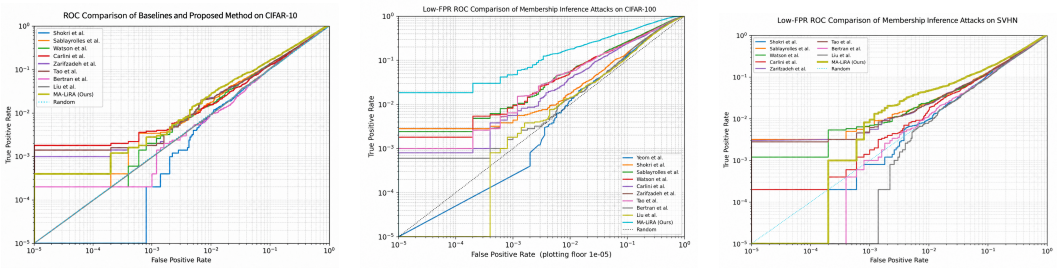
A/E and H/E settings they correspond to only one false positive among 1,000 non-members and are therefore treated as resolution-limited diagnostics.

The log–log curves in Figure 1 reinforce the tabulated low-FPR results: MA-LiRA is generally shifted upward relative to the baselines across all three datasets. The separation is most visible on CIFAR-10 and CIFAR-100, while SVHN remains closer to chance for all attacks.

To test whether the gains of MA-LiRA can be explained simply by a better final-checkpoint score, we evaluate per-example LiRA with several confidence-, logit-, and margin-based endpoint statistics. Even when the best endpoint statistic is selected independently for each metric, MA-LiRA remains substantially stronger across all three datasets (Table 4). The gap is particularly large at low FPRs: on CIFAR-10, TPR@1% FPR increases from 0.0140 to 0.1584, while on CIFAR-100 TPR@0.1% FPR increases from 0.0056 to 0.0472. The improvement also persists on SVHN, indicating that the observed gains cannot be attributed solely to endpoint-score selection.

5 Discussion, Limitations, and Broader Impact

MA-LiRA is a checkpoint-aware privacy audit rather than a final-model-only black-box attack. This stronger access assumption enables the method to exploit training-history information, but it also increases storage and evaluation cost and limits applicability when checkpoints are unavailable or retained irregularly. The compact five-dimensional feature bank reduces redundancy relative to a larger trajectory representation, yet learned fusion can still overfit when the shadow budget is small. Candidate-specific Gaussian calibration also requires sufficient IN/OUT shadow observations and may become unreliable in distributional tails or under substantial target–shadow mismatch.



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

Figure 1: Log–log ROC curves for the principal membership-inference baselines, MA-LiRA, and the random diagonal on CIFAR-10, CIFAR-100, and SVHN. Both axes use identical logarithmic scales, emphasizing the low-FPR regime.

Table 4: Comparison of the strongest endpoint-statistic LiRA configuration with checkpoint-aware MA-LiRA across datasets. For each dataset and metric, the endpoint result is the best value among true-label confidence, log-confidence, unstable/stable logits, true-class output logit, hinge margin, and the corresponding exact clip-free variants. Higher values are better.

Dataset	Method	AUC ↑	TPR@1% FPR ↑	TPR@0.1% FPR ↑
CIFAR-10	Best endpoint statistic	0.5224	0.0140	0.0022
	MA-LiRA (Ours)	**0.7874**	**0.1584**	**0.0282**
CIFAR-100	Best endpoint statistic	0.6380	0.0410	0.0056
	MA-LiRA (Ours)	**0.7984**	**0.1780**	**0.0472**
SVHN	Best endpoint statistic	0.5347	0.0198	0.0028
	MA-LiRA (Ours)	**0.5715**	**0.0474**	**0.0096**

The experiments support the hypothesis that training history contains membership-relevant information beyond the final checkpoint. In particular, the advantage of MA-LiRA persists across datasets, CIFAR-100 classes, difficulty-conditioned audits, and alternative endpoint-statistic choices. These results should nevertheless be interpreted as point estimates: reliable low-FPR evaluation requires sufficiently many independent non-members, and repeated target-model seeds are needed to quantify uncertainty. Moreover, the contribution of any individual memory-aware coordinate cannot be isolated without a matched ablation that holds calibration and fusion fixed.

A stronger membership attack reveals empirical privacy exposure but does not itself provide differential privacy or a defense. We therefore view MA-LiRA primarily as an authorized auditing tool for model developers, internal red teams, and privacy evaluators. Such audits can help identify vulnerable training regimes and evaluate mitigation strategies, while checkpoint access and candidate-level outputs should be appropriately restricted to reduce misuse and unnecessary exposure of sensitive training artifacts.

6 Conclusion

We introduced MA-LiRA, a checkpoint-aware membership-inference attack that uses compact training-trajectory features, candidate-specific shadow calibration, and shadow-only learned fusion. Experiments on CIFAR-10, CIFAR-100, and SVHN show that MA-LiRA improves overall and low-FPR membership inference over the evaluated final-model and trajectory-based baselines under the stated checkpoint-aware access model. Additional endpoint-statistic, class-wise, and difficulty-conditioned analyses further support that the gains are not explained solely by endpoint-score selection or a small subset of examples. Overall, the results indicate that training-history information can provide useful membership evidence beyond a final-checkpoint statistic alone.

Ethics statement

This work is intended solely for authorized privacy auditing. Experiments use public benchmarks and no real sensitive data.

AI use statement

AI tools were used only to improve the manuscript's writing clarity, grammar, and language quality; the authors remain responsible for the final content.

References

Martin Bertran, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu. Scalable membership inference attacks via quantile regression. In *Advances in Neural Information Processing Systems*, volume 36, pages 314–330, 2023.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy*, pages 1897–1914. IEEE, 2022. doi:10.1109/SP46214.2022.9833649.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi:10.1109/CVPR.2016.90.

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

Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership inference attacks by exploiting loss trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2085–2098. ACM, 2022. doi:10.1145/3548606.3560684.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632, 2021.

Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5558–5567. PMLR, 2019.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy*, pages 3–18. IEEE, 2017. doi:10.1109/SP.2017.41.

Jiashu Tao and Reza Shokri. Information-theoretic membership inference for granular quantification of memorization. In *International Conference on Learning Representations*, 2026.

Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*, 2022.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium*, pages 268–282. IEEE, 2018. doi:10.1109/CSF.2018.00027.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58244–58282. PMLR, 2024.