Optimization and Robustness-Informed Membership Inference Attacks for LLMs

Zichen Song [1] **Qixin Zhang** [2] **Ming Li** [3] **Yao Shu*** [4]

Abstract

The proliferation of Large Language Models (LLMs) has raised concerns over training data privacy. Membership Inference Attacks (MIA), aiming to identify whether specific data was used for training, pose significant privacy risks. However, existing MIA methods struggle to address the scale and complexity of modern LLMs. This paper introduces OR-MIA, a novel MIA framework inspired by model optimization and input robustness. First, training data points are expected to exhibit smaller gradient norms due to optimization dynamics. Second, member samples show greater stability, with gradient norms being less sensitive to controlled input perturbations. OR-MIA leverages these principles by perturbing inputs, computing gradient norms, and using them as features for a robust classifier to distinguish members from non-members. Evaluations on LLMs (70M to 6B parameters) and various datasets demonstrate that OR-MIA outperforms existing methods, achieving over 90% accuracy. Our findings highlight a critical vulnerability in LLMs and underscore the need for improved privacy-preserving training paradigms.

1. Introduction

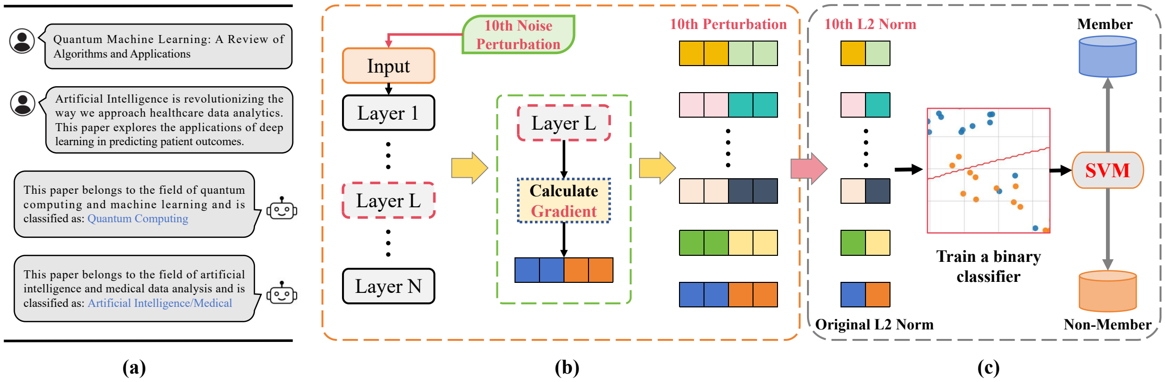
The rapid advancement of Large Language Models (LLMs) has undeniably revolutionized numerous fields, offering unprecedented capabilities in natural language understanding, generation, and reasoning. These models, underpinning applications from sophisticated chatbots to automated code generation, derive their power from being trained on vast and diverse datasets. Yet, this reliance on extensive training data, which often includes sensitive personal, proprietary, or copyrighted information, casts a significant shadow: the potential for privacy breaches. As LLMs become more integrated into our digital lives, understanding and mitigating their privacy vulnerabilities is paramount.

Among the most direct and concerning privacy threats are Membership Inference Attacks (MIA). An MIA aims to determine whether a specific data point was part of the training set of a target model. A successful MIA can have severe consequences, ranging from exposing sensitive individual records used in training a healthcare LLM to revealing confidential corporate documents or proprietary code. Such breaches not only violate individual privacy and intellectual property rights but also erode public trust in AI systems, potentially hindering their adoption and societal benefit. The development of robust MIA, therefore, serves a dual purpose: it highlights existing vulnerabilities that need addressing and provides a crucial benchmark for evaluating the efficacy of privacy-preserving techniques, guiding future research in this area (Biderman et al., 2023a;b; Andonian et al., 2023; Bertran et al., 2023).

Despite extensive research on MIA in traditional machine learning, their application to modern LLMs presents unique and formidable challenges. The sheer scale of LLM parameters, the massive size of their training corpora, and the complexity of their architectures often render conventional MIA techniques ineffective. For instance, methods based on loss functions (Yeom et al., 2018), while foundational, tend to perform poorly on large-scale datasets typical of LLMs. Similarly, approaches leveraging model compression or intermediate feature extraction (Carlini et al., 2021), or those utilizing prediction uncertainty (Shi et al., 2023), have shown improvements but still struggle significantly when applied to LLMs, often failing to achieve success rates substantially better than random guessing (Carlini et al., 2019; 2021; 2023). Even more recent LLM-specific MIA, such as those employing model perturbations like MoPe (Li et al., 2023) or self-prompt calibration techniques for fine-tuned models like SPV-MIA (Fu et al., 2024), face limitations in generalizability and consistent success across diverse and complex LLM structures. Consequently, many existing MIA strategies, including LOSS, Zlib, and Min-K,

[1] Department of Computer Science, Lanzhou University, China [2] College of Computing and Data Science, Nanyang Technological University, Singapore [3] Guangming Laboratory, China [4] School of Information Science, The Hong Kong University of Science and Technology (Guangzhou), China. Correspondence to: Yao shu <yaoshu@hkust-gz.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).



Quantum Machine Learning: A Review of Algorithms and Applications

Artificial Intelligence is revolutionizing the way we approach healthcare data analytics. This paper explores the applications of deep learning in predicting patient outcomes.

This paper belongs to the field of quantum computing and machine learning and is classified as: Quantum Computing

This paper belongs to the field of artificial intelligence and medical data analysis and is classified as: Artificial Intelligence/Medical

Figure 1. Framework of our proposed membership inference attack method OR-MIA. (a) Examples of inference tasks for LLMs. (b) Perturbation via Gaussian noise to generate perturbed inputs. (c) Gradient norm analysis and SVM-based classification to distinguish members from non-members.

alongside these newer approaches, have notable limitations and are often insufficient to pose a substantial threat to large pre-trained LLMs (Black et al., 2021; Brown et al., 2020; Carlini et al., 2022).

Motivated by these limitations, this paper introduces OR-MIA, a novel and theoretically-grounded MIA framework designed to significantly enhance attack efficacy against LLMs. Our method, visually outlined in Figure 1, leverages two fundamental insights related to model optimization and input robustness: first, data points seen during training (Figure 1(a)) are expected to yield smaller gradient norms due to the model **optimization state**; and second, these member samples tend to exhibit more stable gradient norms under controlled input perturbations due to greater **input robustness** compared to non-members. OR-MIA operationalizes these principles by systematically generating a series of perturbed versions for each target input sample. For both the original and each perturbed input, we compute the L2 norm of the gradients of the model loss function (Figure 1(b)). This sequence of gradient norms, which captures both the initial optimization state and the stability of this state under perturbation, forms a distinctive feature vector. This vector is then fed into a supervised classifier, such as a Support Vector Machine (SVM), to infer the membership status of a sample (Figure 1(c)).

Our contributions are threefold:

- **A Novel Optimization and Robustness-Informed MIA Methodology**: We introduce OR-MIA, a theoretically-grounded attack framework that uniquely synergizes insights from model optimization and input robustness. By systematically analyzing the L2 norm of gradients under a sequence of controlled input perturbations, OR-MIA constructs a highly discriminative feature signature to identify training data membership, moving beyond traditional reliance on static output statistics.
- **Extensive Empirical Validation and Superior Performance**: We conduct comprehensive experiments across a diverse suite of Large Language Models, ranging from 70M to 6B parameters, and on multiple real-world datasets. Our results consistently demonstrate that OR-MIA significantly outperforms existing state-of-the-art membership inference attacks, achieving substantial improvements in both attack accuracy and sample efficiency, thereby establishing a new, more potent benchmark for LLM privacy evaluation.
- **In-depth Ablation Studies Yielding Novel Insights**: Through rigorous ablation studies, we meticulously dissect the components of OR-MIA, empirically validating our core hypotheses regarding the individual and combined efficacy of gradient norm analysis and perturbation-based robustness signals. These studies not only confirm the mechanisms behind the success of OR-MIA but also provide novel insights into LLM vulnerabilities, such as layer-specific sensitivities to membership signals, deepening the understanding of how and where privacy leakage manifests in these complex architectures.

2. Problem Setup

Membership Inference Attack (MIA), which poses a significant privacy threat to machine learning models, aims to determine whether a given data sample x was part of the training dataset D_{train} used to train a target model f_θ, without direct access to $\mathcal{D}_{\text{train}}$. In the context of LLMs, the data sample $x = (x_1, x_2, \ldots, x_T)$ is a sequence of tokens representing text or code, the label y is the next token to be predicted, and f_θ is an auto-regressive language model that outputs a predicted probability of the next token given a prefix, denoted as $p(x_t|x_1, \ldots, x_{t-1}; f_\theta)$. Formally, the MIA problem is a binary classification task: Given a data sample x and access to the target LLM f_θ, the attacker aims to predict a membership label z where $z = 1$ indicates x is a member of $\mathcal{D}_{\text{train}}$ and $z = 0$ indicates it is not. Typically, the attacker constructs a MIA score function $s(x, f_\theta)$ based on the behavior of the target LLM f_θ to predict membership:

$$\hat{z} = \mathbb{I}(s(x, f_\theta) < \tau) \,. \tag{1}$$

Here, τ is a decision threshold, which is optimized on a separate validation set, D_{val}, to minimize the MIA error rate. Obviously, the design of the function $s(x, f_\theta)$ and the selection of the threshold τ are critical for the efficacy of an MIA. Of note, in this setting, we are assumed to have the access to the parameter θ and the corresponding gradient of the target LLM f_θ, resulting from the increasing prevalence of open-sourced LLMs.

Despite recent advancements on MIA, existing methods like (Yeom et al., 2018; Carlini et al., 2021; Shi et al., 2023; Li et al., 2023; Fu et al., 2024) often overlook the direct link between a sample and the optimization state of the target model, or how model robustness to input variations differs for members versus non-members. Our approach, OR-MIA, is designed to leverage these underexplored signals.

3. Optimization and Robustness-Informed Membership Inference Attack

Our approach, OR-MIA (see Figure 1), is built upon two complementary principles designed to probe the relationship between a data sample and a trained LLM. We first take the gradient norm as a direct indicator of the model optimization state with respect to the sample (Sec. 3.1). We then assess the model robustness to input perturbations, hypothesizing that training samples exhibit more stable behavior (Sec. 3.2). These principles guide the design of our membership inference algorithm (Sec. 3.3).

3.1. Gradient Norm as a Probe for Optimization State

Our first principle leverages the gradient norm, $\|\nabla_\theta \ell(f_\theta(x), y)\|_2$, as a direct probe into the optimization state of the model f_θ concerning a specific input x. This offers a more fundamental and dynamic signal compared to static model outputs like loss or probabilities. While outputs merely reflect the final fit and can be misleadingly similar for members and well-generalized non-members (making them inherently ambiguous), the gradient norm captures the *dynamics* of optimization. Specifically, it quantifies the model sensitivity to x and the magnitude of the parameter update that would be driven by that sample, revealing its potential influence on θ.

This sensitivity is intrinsically tied to the training history. Gradient-based optimization methods (e.g., SGD, AdamW (Loshchilov & Hutter, 2019)), which underpin most LLM training, iteratively adjust parameters θ to minimize a loss function ℓ. As formalized in Theorem 3.1, a necessary condition for reaching a local minimum θ^* is that the gradient vanishes, $\nabla_\theta \ell(\theta^*) = 0$.

Theorem 3.1. *Let $\ell(\theta)$ be a continuously differentiable function. A point θ^* is a local minimizer of ℓ if and only if:*

$$\nabla_\theta \ell(\theta)\big|_{\theta=\theta^*} = 0 \,. \tag{2}$$

For a sample (x, y) present in the training set ($\mathcal{D}_{\text{train}}$), the optimization process directly aims to reduce $\ell(f_\theta(x), y)$. Consequently, for a well-trained model, the parameters θ have been specifically adjusted to accommodate these training samples. A direct consequence is that the gradient norm computed for these samples is expected to be significantly smaller than for unseen samples, approaching the zero condition dictated by optimality:

$$\begin{aligned} \left\|\nabla_\theta \ell(f_\theta(x), y)\right\|_2 &\approx 0 \quad \text{for } x \in \mathcal{D}_{\text{train}} \,, \\ \left\|\nabla_\theta \ell(f_\theta(x), y)\right\|_2 &> \epsilon \quad \text{for } x \notin \mathcal{D}_{\text{train}} \,. \end{aligned} \tag{3}$$

Therefore, the L2 norm of the model gradients serves as a theoretically-backed measure reflecting whether a sample actively shaped the model parameters during training. This provides our first key signal for distinguishing members from non-members.

3.2. Perturbation as a Probe for Input Robustness

Our second principle leverages input robustness as a distinguishing characteristic between training members and non-members. While LLMs trained on vast datasets achieve remarkable generalization, the nature of their learned representations and decision boundaries can differ significantly for data seen during training versus unseen data. We are motivated by manifold learning, i.e., machine learning models, including LLMs, implicitly learn a lower-dimensional manifold on which the training data predominantly lies (Gorban & Tyukin, 2018; Narayanan & Mitter, 2010). We hypothesize that member samples, having directly shaped this learned manifold and driven the optimization process, reside in regions of the input space where the model behavior is inherently more stable and its decision function locally smoother (Neyshabur et al., 2015). Consequently, when subjected to controlled input perturbations, the model sensitivity to its parameters, as quantified by the gradient norm, should exhibit greater stability for these member samples. Small perturbations are likely to keep members within or near these well-characterized regions of the learned manifold, leading to consistent gradient responses. In contrast, non-member samples, which may lie further from this core manifold, in sparser regions, or near more complex or irregular parts of the decision boundary, are anticipated to elicit more volatile and erratic gradient norm responses under similar perturbations, as the model navigates less familiar or less smoothly defined input territories.

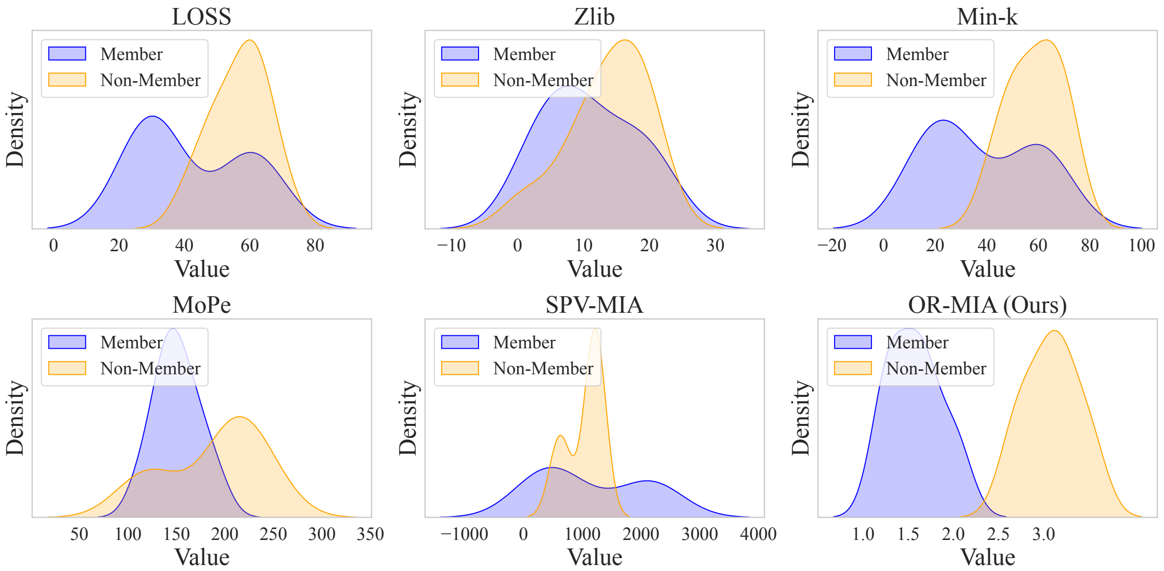


Figure 2. Kernel Density Estimation (KDE) plots of decision scores for members (blue) and non-members (orange) for various MIA methods on The Pile dataset using 3B. OR-MIA demonstrates significantly better separation between member and non-member distributions compared to LOSS, Zlib, and Min-k, indicating a more effective decision boundary.

To operationalize this principle, we introduce controlled perturbations and observe the resulting gradient norm changes:

(a) Generate Perturbed Inputs. For a given input sample x, we create a sequence of N slightly modified versions x_i ($i = 1, \dots, N$). This is achieved by adding Gaussian noise $\xi_i \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ with increasing scale: $x_i \leftarrow x + (i-1)\xi_i$. In our experiments, we use $N = 10$ and $\sigma = 0.1$ (see Appendix A.1). Note that x_1 corresponds to the original, unperturbed input ($i = 1$, scale is 0).

(b) Compute Perturbed Gradient Norms. For each perturbed sample x_i, we compute the L2 norm of the gradients with respect to the model parameters, using the original label or target sequence y: $g_i = \|\nabla_\theta \ell(f_\theta(x_i), y)\|_2$. This yields a set of N gradient norms, $G = \{g_1, g_2, \dots, g_N\}$.

This set G captures the change of gradient norms under increasing input perturbation. According to our robustness hypothesis, for member samples, we expect the values in G (particularly $g_2, \dots, g_N$ relative to g_1) to exhibit less variance or smaller magnitude changes compared to non-member samples, which we will validate in our ablation study in Appendix A.6. This stability profile therefore provides our second key signal.

3.3. Final Algorithm: Synergistic Classification

The OR-MIA framework synthesizes the insights derived from both the optimization state (Section 3.1) and input robustness (Section 3.2) into a unified membership inference procedure. The core of the algorithm involves constructing a feature representation based on the gradient norm dynamics under perturbation and employing a supervised classifier to distinguish between members and non-members.

Specifically, for each x, we first compute the sequence of N gradient norms $G = \{g_1, g_2, \dots, g_N\}$ in Section 3.2. The gradients used to compute these norms are derived from a specific model layer (or set of layers) determined empirically to maximize discriminability, typically selected from the middle layers based on performance on a calibration dataset, as informed by our ablation studies (Section A.6). This sequence forms the basis of our feature vector $\mathbf{v}_x = (g_1, g_2, \dots, g_N) \in \mathbb{R}^N$. This vector $\mathbf{v}_x$ captures both the initial optimization state w.r.t. x (g_1) and the stability of gradient norm under increasing input perturbations ($g_2, \dots, g_N$). We then frame MIA as a binary classification problem, where the objective is to predict the membership (member vs. non-member) of a sample x based on its feature representation $\mathbf{v}_x$. For this purpose, we apply Support Vector Machine (SVM), a well-established algorithm effective for high-dimensional classification tasks. To accommodate potentially complex and non-linear decision boundaries in the feature space defined by $\mathbf{v}_x$, we utilize the kernel trick within the SVM framework. Specifically, we employ the RBF kernel:

$$k(\mathbf{v}_i, \mathbf{v}_j) = \exp\left(-\gamma \|\mathbf{v}_i - \mathbf{v}_j\|_2^2\right)\ , \tag{4}$$

where γ is a kernel parameter. The RBF kernel allows the SVM to implicitly map the feature vectors into a higher-dimensional space, enabling the identification of non-linear separation patterns between member and non-member representations.

The SVM is trained on a dedicated calibration dataset containing samples with known membership status (i.e., con-

Table 1. MIA success rates across datasets and model sizes. Our OR-MIA consistently achieves the highest accuracy, demonstrating the effectiveness of combining optimization state and robustness signals. Sample size for calibration/testing is 30 per class per run. We use **bold** and $\boxed{\text{boxed}}$ value to denote the best and second best performance on each dataset while under the same model size.

Dataset	Method	Model Size				
		70M	160M	1B	3B	6B
Arxiv	LOSS	0.531 ± 0.069	0.544 ± 0.039	0.561 ± 0.038	0.573 ± 0.042	0.587 ± 0.084
	Zlib	0.543 ± 0.068	0.544 ± 0.061	0.608 ± 0.062	0.609 ± 0.059	0.631 ± 0.064
	Min-k	0.528 ± 0.064	0.537 ± 0.072	0.562 ± 0.058	0.556 ± 0.051	0.586 ± 0.063
	MoPe	$\boxed{0.724}$ ± 0.078	$\boxed{0.720}$ ± 0.043	$\boxed{0.734}$ ± 0.044	0.740 ± 0.048	0.624 ± 0.059
	SPV-MIA	0.531 ± 0.062	0.543 ± 0.071	0.561 ± 0.060	0.563 ± 0.048	0.589 ± 0.062
	OR-MIA (ours)	**0.927** ± 0.068	**0.941** ± 0.058	**0.933** ± 0.047	**0.933** ± 0.052	**0.874** ± 0.048
	↪ w/o Perturb.	0.663 ± 0.058	0.669 ± 0.062	0.692 ± 0.049	$\boxed{0.779}$ ± 0.030	$\boxed{0.812}$ ± 0.061
HackerNews	LOSS	0.544 ± 0.067	0.569 ± 0.033	0.591 ± 0.074	0.612 ± 0.029	0.623 ± 0.059
	Zlib	0.553 ± 0.059	0.588 ± 0.038	0.617 ± 0.042	0.618 ± 0.048	0.634 ± 0.040
	Min-k	0.558 ± 0.038	0.556 ± 0.059	0.607 ± 0.037	0.627 ± 0.018	0.622 ± 0.051
	MoPe	$\boxed{0.694}$ ± 0.092	$\boxed{0.714}$ ± 0.039	$\boxed{0.735}$ ± 0.028	0.728 ± 0.033	0.612 ± 0.063
	SPV-MIA	0.557 ± 0.036	0.555 ± 0.060	0.609 ± 0.039	0.627 ± 0.019	0.624 ± 0.052
	OR-MIA (ours)	**0.921** ± 0.069	**0.943** ± 0.053	**0.944** ± 0.058	**0.940** ± 0.039	**0.979** ± 0.019
	↪ w/o Perturb.	0.642 ± 0.048	0.691 ± 0.039	0.712 ± 0.044	$\boxed{0.752}$ ± 0.043	$\boxed{0.794}$ ± 0.039
The Pile	LOSS	0.514 ± 0.076	0.538 ± 0.075	0.569 ± 0.070	0.581 ± 0.051	0.591 ± 0.068
	Zlib	0.522 ± 0.029	0.544 ± 0.073	0.569 ± 0.047	0.609 ± 0.041	0.615 ± 0.044
	Min-k	0.531 ± 0.069	0.554 ± 0.043	0.559 ± 0.067	0.613 ± 0.046	0.591 ± 0.070
	MoPe	$\boxed{0.713}$ ± 0.063	$\boxed{0.734}$ ± 0.041	$\boxed{0.739}$ ± 0.031	0.746 ± 0.052	0.612 ± 0.052
	SPV-MIA	0.532 ± 0.071	0.551 ± 0.038	0.561 ± 0.069	0.607 ± 0.047	0.591 ± 0.066
	OR-MIA (ours)	**0.957** ± 0.032	**0.951** ± 0.038	**0.950** ± 0.048	**0.960** ± 0.031	**0.946** ± 0.041
	↪ w/o Perturb.	0.671 ± 0.052	0.681 ± 0.039	0.699 ± 0.038	$\boxed{0.781}$ ± 0.021	$\boxed{0.778}$ ± 0.028

firmed members from $\mathcal{D}_{\text{train}}$ and confirmed non-members). During training, the SVM learns the optimal separating hyperplane (in the kernel-induced feature space) that maximizes the margin between the two classes. Once trained, the SVM classifier assigns a membership score to a new sample x based on its feature vector $\mathbf{v}_x$. This score is from the decision function of SVM, which typically corresponds to the signed distance of the sample feature vector from the learned hyperplane in the transformed space. A higher score indicates a higher likelihood that the sample x was part of the original training dataset $\mathcal{D}_{\text{train}}$. This final score effectively integrates the information from both the initial gradient magnitude and its stability profile under perturbation, as optimally weighted by the trained SVM classifier.

4. Experiments

In this section, we empirically validate the effectiveness of OR-MIA, our proposed optimization and robustness-informed membership inference attack. We conduct a comprehensive evaluation across diverse datasets and LLM architectures of varying scales. Our experiments are designed to: (1) demonstrate the superior performance of OR-MIA compared to state-of-the-art MIA baselines, (2) analyze the distribution separation achieved by different methods, highlighting the discriminative power of our approach, (3) assess the sample efficiency of OR-MIA, and (4) perform ablation studies to dissect the contributions of the core components of our method, namely the gradient norm signal and the perturbation-based robustness analysis. Experimental setup is detailed in Appendix A.2.

4.1. Main Results

Superior Attack Accuracy. Table 1 presents the primary MIA success rates across all datasets and model sizes. OR-MIA consistently and significantly outperforms all baseline methods, often achieving accuracy above 0.90, approaching near-perfect inference in several settings (e.g., 0.979 on HackerNews with LLaMa-6B, 0.960 on The Pile with LLaMa-3B). This substantial improvement underscores the efficacy of our core principles. Baselines relying solely on output statistics such as LOSS, Zlib, or Min-k struggle, especially on larger models where generalization might obscure membership signals in outputs alone. Methods like MoPe and SPV-MIA show improvements but are still considerably outperformed by OR-MIA. Crucially, OR-MIA also demonstrates a large margin over the "No Perturbation" baseline (e.g., 0.933 vs 0.779 on Arxiv 3B, 0.979 vs 0.794 on HackerNews 6B). This directly validates the hypothesis that incorporating the stability of gradient norm under perturbation (Section 3.2) provides a powerful, complementary signal to the initial gradient norm magnitude (Section 3.1).

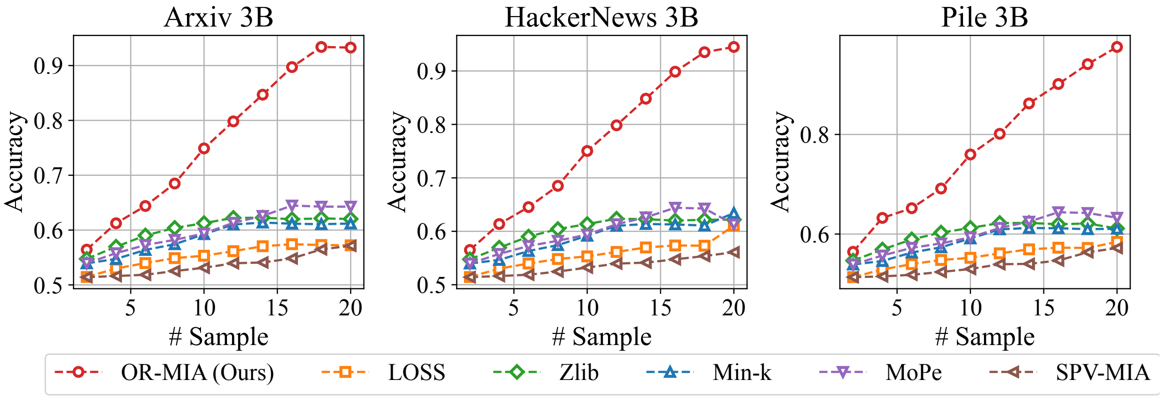


Figure 3. Comparison of sample-size efficiency on LLaMa-3B. OR-MIA achieves high accuracy with significantly fewer calibration samples compared to what would be expected from lower-performing baselines.

The feature vector $\mathbf{v}_x$ capturing both optimization state and robustness, combined with the non-linear classification power of the RBF-SVM (Section 3.3), proves highly effective at distinguishing members from non-members.

Enhanced Distribution Separation. To gain deeper insight into why OR-MIA performs better, we visualize the distributions of the decision scores (or underlying features) for members and non-members. Figure 2 shows the Kernel Density Estimation (KDE) plots of the final SVM decision scores for various methods on The Pile dataset. OR-MIA exhibits a remarkably clear separation between the member (blue) and non-member (orange) distributions, with minimal overlap. This contrasts sharply with baselines like LOSS, Zlib, and Min-k, which show significant overlap, indicating poor discriminative power. This superior separation arises directly from our methodology. As hypothesized in Section 3.1, members tend to have lower initial gradient norms g_1, pushing their feature vectors $\mathbf{v}_x$ in one direction. Furthermore, as validated by Figure 2, members exhibit significantly smaller changes in gradient norm under perturbation compared to non-members. This robustness signal (captured by $\{g_2, \cdots, g_N\}$) provides an orthogonal dimension for separation. The T-SNE visualization in appendix Figure 6 vividly illustrates this: without perturbation (left), member and non-member clusters show considerable mixing; with perturbation (right, representing the feature space used by OR-MIA), the clusters become much more distinct and separable. The RBF-SVM effectively learns a non-linear boundary in this enhanced feature space (Equation 4) to exploit both the lower magnitude and higher stability of gradient norms for members, resulting in the clean separation observed in the KDE plots.

High Sample Efficiency. In practical MIA scenarios, the attacker may only have access to a limited number of samples with known membership status for calibration. We evaluate the sample efficiency by varying the number of calibration samples used to train the SVM classifier. Figure 3 shows the attack accuracy as a function of the calibration set size for the LLaMa-3B model across the three datasets. OR-MIA demonstrates remarkable sample efficiency. It achieves high accuracy (often > 0.70) with as few as 10-20 known samples per class. This efficiency stems from the strong and informative nature of the feature vector $\mathbf{v}_x$. The initial gradient norm g_1 provides a direct, potent signal related to the optimization state (Section 3.1), which is effective even with few examples. The subsequent components $\{g_2, \cdots, g_N\}$ add robustness information (Section 3.2), making the overall signal statistically reliable and allowing the SVM to learn an effective decision boundary quickly, even from limited data. This makes OR-MIA particularly potent in realistic, data-constrained attack scenarios.

5. Conclusion

In this work, we presented a novel membership inference attack methodology for large language models (LLMs), combining gradient norm analysis and perturbation confidence to significantly enhance attack success rates. Our findings demonstrate that this unified framework effectively addresses the limitations of traditional methods, achieving robust performance across diverse datasets and model scales. This work highlights the urgent need for re-evaluating privacy-preserving strategies in LLMs, with implications for both theoretical advancements and practical defenses. We hope this study inspires further research into balancing model performance and privacy in the era of large-scale AI systems.

References

Andonian, A., Anthony, Q., Biderman, S., Black, S., Gali, P., and Gao, L. Gpt-neox: Large scale autoregressive language modeling in pytorch. 2023. URL https://www.github.com/eleutherai/gpt-neox.

Bertran, M. A., Tang, S., Roth, A., Kearns, M., Morgenstern, J. H., and Wu, S. Scalable membership inference attacks via quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, 2023. NeurIPS.

Biderman, S., Prashanth, U. S., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raf, E. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, 2023a. NeurIPS.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the International Conference on Learning Representations (ICLR 2023)*, Riyadh, Saudi Arabia, 2023b. ICLR.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. In *Proceedings of the 2021 Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2021. NeurIPS. URL http://github.com/eleutherai/gpt-neo.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, 2020. NeurIPS.

Carlini, N., Liu, C., Erlingsson, U., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of the USENIX Security Symposium*, Santa Clara, CA, 2019. USENIX.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., and Erlingsson, U. Extracting training data from large language models. In *Proceedings of the USENIX Security Symposium*, Vancouver, Canada, 2021. USENIX.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *Proceedings of the IEEE Symposium on Security and Privacy*, San Francisco, CA, 2022. IEEE.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *Proceedings of the International Conference on Learning Representations (ICLR 2023)*, Riyadh, Saudi Arabia, 2023. ICLR.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. Do membership inference attacks work on large language models? *ArXiv*, abs/2402.07841, 2024. URL https://arxiv.org/abs/2402.07841.

Fu, W., Wang, H., Gao, C., Liu, G., Li, Y., and Jiang, T. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration, 2024. URL https://arxiv.org/abs/2311.06062.

Gorban, A. N. and Tyukin, I. Y. Blessing of dimensionality: mathematical foundations of the statistical physics of data. *CoRR*, abs/1801.03421, 2018.

Li, M., Wang, J., Wang, J., and Neel, S. MoPe: Model perturbation based privacy attacks on language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13647–13660, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.842. URL https://aclanthology.org/2023.emnlp-main.842/.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*. ICLR, 2019.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., and Zanella-Beguelin, S. Analyzing leakage of personally identifiable information in language models. In *Proceedings of the IEEE Symposium on Security and Privacy*, San Francisco, CA, 2023. IEEE.

Magnusson, I., Bhagia, A., Hofmann, V., Soldaini, L., Jha, A. H., Tafjord, O., Schwenk, D., Walsh, E. P., Elazar, Y., Lo, K., Groenveld, D., Beltagy, I., Hajishirzi, H., Smith, N. A., Richardson, K., and Dodge, J. Paloma: A benchmark for evaluating language model fit. Technical report, Allen Institute for AI, 2023. URL https://paloma.allen.ai/.

Mattern, J., Mireshghallah, F., Jin, Z., Schoelkopf, B., Sachan, M., and Berg-Kirkpatrick, T. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics (ACL)*, Toronto, Canada, 2023. ACL.

Meeus, M., Jain, S., Rei, M., and de Montjoye, Y.-A. Did the neurons read your book? document-level membership inference for large language models. *arXiv* *preprint arXiv:2310.15007*, 2023. URL https://arxiv.org/abs/2310.15007.

Min, S., Gururangan, S., Wallace, E., Hajishirzi, H., Smith, N. A., and Zettlemoyer, L. Silo language models: Isolating legal risk in a nonparametric datastore. In *Regulatable ML Workshop at NeurIPS*, New Orleans, LA, 2023. NeurIPS.

Mireshghallah, F., Uniyal, A., Wang, T., Evans, D., and Berg-Kirkpatrick, T. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Online, 2022. EMNLP.

Narayanan, H. and Mitter, S. K. Sample complexity of testing the manifold hypothesis. In *NIPS*, pp. 1786–1794. Curran Associates, Inc., 2010.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015.

Puerto, H., Gubri, M., Yun, S., and Oh, S. J. Scaling up membership inference: When and how attacks succeed on large language models. *ArXiv*, 2411.00154, 2024. doi: 10.48550/arXiv.2411.00154. URL https://arxiv.org/abs/2411.00154.

Raganato, A. and Tiedemann, J. An analysis of encoder representations in transformer-based machine translation. In *BlackboxNLP@EMNLP*, pp. 287–297. Association for Computational Linguistics, 2018.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *Regulatable ML Workshop at NeurIPS 2023*, New Orleans, LA, 2023. NeurIPS.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *Proceedings of the IEEE Symposium on Security and Privacy*, San Jose, CA, 2017. IEEE.

Skean, O., Arefin, M. R., LeCun, Y., and Shwartz-Ziv, R. Does representation matter? exploring intermediate layers in large language models. *arXiv preprint arXiv:2412.09563*, 2024.

Wang, X., Wu, L., and Guan, Z. Graddiff: Gradient-based membership inference attacks against federated distillation with differential comparison. *Information Sciences*, 658:120068, 2024. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2023.120068.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proceedings of the 31st IEEE Computer Security Foundations Symposium (CSF 2018)*, Cambridge, MA, 2018. IEEE.

A. Appendix

A.1. Related Work

Traditional MIA and their Limitations with LLMs. The foundational framework for membership inference attacks (MIA) utilizes multiple shadow models to simulate target model behavior and extract data for the attack (Shokri et al., 2017). This approach, however, assumes attacker access to datasets with distributions identical to the target model training data, a condition rarely met in practice. It has been shown that while traditional MIA perform well on smaller models, they falter against large language models (LLMs), particularly in short sequence analysis where member/non-member distinctions blur (Duan et al., 2024). Traditional MIA, focusing on model fit to training data, overlook challenges posed by LLMs' large-scale datasets, often performing near chance levels. This underscores the need to reassess how model memorization of training data is evaluated in LLMs (Lukas et al., 2023; Magnusson et al., 2023; Mattern et al., 2023).

Gradient-based MIA and their Challenges in the LLM Context. Gradient-based MIA have been proposed, achieving attacks by comparing gradient differences of jointly distilled models (Carlini et al., 2021). While theoretically promising, their practical application to LLMs has been constrained by the massive scale and complexity of modern models. More recently, GradDiff, a passive gradient-based MIA, was introduced to federated learning and federated distillation (Wang et al., 2024). Despite some improvements, GradDiff struggles with the intricacies of LLMs, including their highly structured and diverse input data. Gradient-based methods often fail to capture the nuanced model behavior with respect to high-dimensional natural language. The vast parameter spaces and sophisticated optimization strategies in LLMs further complicate these methods, diminishing the directness and effectiveness of gradient information for inferring specific data points (Meeus et al., 2023; Min et al., 2023; Mireshghallah et al., 2022).

Emerging LLM-Specific MIA and their Current Frontiers. Recent works have developed new MIA methods specifically targeting LLMs. E.g., MoPe uses perturbations to explore privacy leakage but faces limitations with complex LLM structures (Li et al., 2023). SPV-MIA enhances attack effectiveness on fine-tuned LLMs through self-prompt calibration (Fu et al., 2024). However, challenges persist, particularly with very large-scale LLMs and fine-tuning on diverse task-specific datasets (Black et al., 2021; Brown et al., 2020; Carlini et al., 2022). Concurrently, the aggregation of MIA scores across multiple documents has been demonstrated to successfully perform MIA at the dataset level, offering a practical technique for broader inference (Puerto et al., 2024). While these approaches advance MIA research, they still encounter scalability issues and highlight the need for further refinement to handle the complexity of modern LLMs.

Addressing these limitations, this paper introduces a novel gradient-based MIA enhanced by perturbation confidence. We systematically perturb samples multiple times to amplify prediction uncertainty, thereby exposing LLM sensitivity to such changes. This approach aims to provide more reliable insights into model memorization and significantly improve attack efficacy against large-scale LLMs.

A.2. Experimental Setup

Datasets and Models. We evaluate MIA performance on three widely used text corpora: **Arxiv** (scientific papers), **HackerNews** (online discussions), and **The Pile** (a diverse large-scale dataset). We target LLMs from the Pythia suite (Biderman et al., 2023b) with varying parameter counts (70M, 160M, 1B) and LLaMa models (LLaMa-3B, LLaMa-6B), representing a range of model capacities and architectures prevalent in open-source LLMs.

Evaluation Metric. We report the standard MIA success rate (Accuracy), defined as the proportion of samples correctly classified as either member or non-member. Results are averaged over multiple runs (e.g., 5 runs), and we report the mean ± standard deviation. A random guess corresponds to an accuracy of 0.5.

Implementation Details. For OR-MIA, we generate $N = 10$ perturbed samples per input using Gaussian noise with $\sigma = 0.1$, as detailed in Section 3.2. The feature vector is computed using the L2 norm of gradients w.r.t. all model parameters. We train a Support Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel (Equation 4) on a balanced calibration set of known members and non-members (e.g., 1000 samples each). The number of test sets is approximately between 10 and 100. Hyperparameters for the SVM are tuned using cross-validation on the calibration set.

A.3. Hyperparameters

The following hyperparameters were used in the training and evaluation of the model:

- **Random Seed** (`set_seed`): The random seed was set to `42` to ensure reproducibility of the experiments across different runs.
- **Model Name** (`model_name`): The pre-trained model `stablelm-base-alpha-3b-v2` was used for the fine-tuning experiments.
- **Maximum Sequence Length** (`max_length`): The maximum length for the tokenized input sequences was set to `128`, ensuring that input text sequences longer

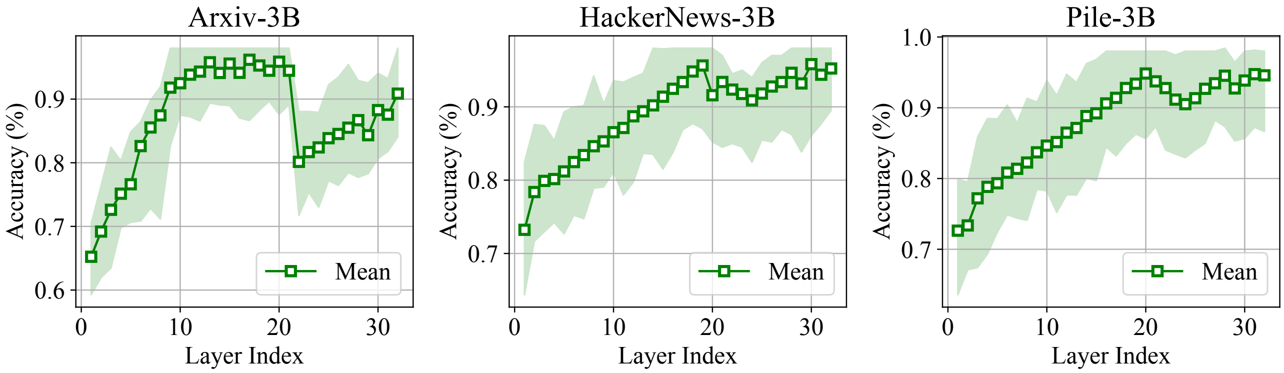


Figure 4. Layer-specific vulnerability to OR-MIA on Arxiv-3B. Attack accuracy is plotted as a function of the LLM layer from which gradients are derived. The results demonstrate a non-monotonic relationship, with middle layers (e.g., 15th layer) exhibiting the highest vulnerability, compared to shallower (e.g., 5th layer) or deeper (e.g., 30th layer) layers. Shaded areas indicate the max-min range over runs.

than this value are truncated and shorter sequences are padded accordingly.

- **Batch Size** (`batch_size`): A batch size of `8` was used for both training and evaluation to balance between memory usage and computational efficiency.

- **Learning Rate** (`lr`): The learning rate for the Adam optimizer was set to `1e-5`, providing a fine-grained adjustment of model weights during training.

- **Number of Epochs** (`epochs`): The model was trained for `1` epoch. This number was chosen to observe the initial learning dynamics without overfitting on a small dataset.

- **Layer Freezing** (`freeze_all_layers`): Initially, all layers of the model were frozen to prevent weight updates, and only specific layers were subsequently unfrozen for training.

- **Layer Unfreezing** (`unfreeze_layer`): Specific layers of the model were selectively unfrozen during the evaluation process to assess their individual contribution to model performance.

A.4. Justification for Fine-tuning on Preset Membership Dataset

In this study, we fine-tune the model on a dataset of "preset members" to ensure the model can properly identify whether the data is part of the training set that the large language model (LLM) was exposed to. This procedure is essential for the following reasons:

Large Language Models May Not Contain Our "Member" Data: When we load a pre-trained model from an open-source platform, we are working with a model that has been trained on a broad, general dataset, such as those collected from web sources or public text. However, the data we are interested in—referred to as "member" data—may not be part of this pre-training dataset. This is because the open-source model may have never been trained on the specific dataset we are using. Consequently, without fine-tuning, the model may not be capable of recognizing or differentiating the "member" data, since it was never exposed to this data during its original training.

The Role of Fine-tuning: Fine-tuning on a dataset containing our "member" data allows the model to adjust its weights to recognize patterns specific to this dataset. Even though the pre-trained model has learned general language patterns, fine-tuning enables the model to specialize in identifying whether a given input comes from the "member" dataset or not. Without this fine-tuning step, the model will lack the necessary internal representation to make accurate membership determinations, as it was not trained on the specific data that we consider as "members."

In summary, fine-tuning is essential to ensure the model can distinguish between data that belongs to the training set it was originally exposed to (i.e., "members") and data it has never seen before. Open-source pre-trained models are not guaranteed to have seen our specific data during their original training process, so fine-tuning on a labeled dataset is the only way to allow the model to identify membership reliably.

A.5. Limitation

Despite the promising results presented by Normia, there are several limitations that should be acknowledged. First, while our method significantly enhances Membership Inference Attacks (MIA) on LLMs, it relies heavily on the availability of a fully trained model, which may not always

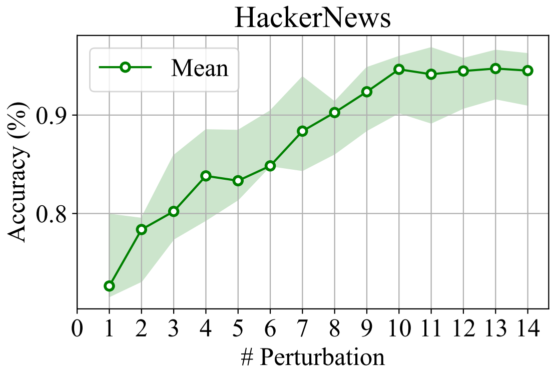
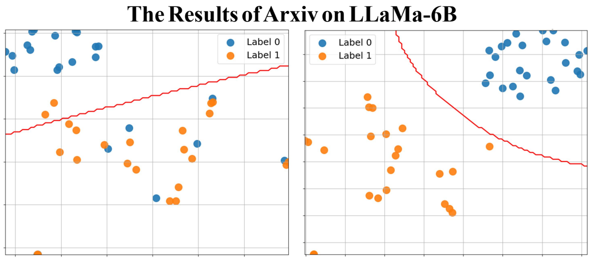


Figure 5. Impact of number of perturbation on OR-MIA accuracy for HackerNews. Attack accuracy improves with increasing N, saturating around $N = 10$, validating the chosen number of perturbations. Shaded area indicates max-min range over runs.



(a) No perturbation (b) perturbation

Figure 6. T-SNE of feature representations for members (blue) and non-members (orange) on Arxiv-6B. Left: OR-MIA w/o perturbation (using only g_1). Right: OR-MIA (using $\{g_1, \cdots, g_N\}$). The feature vector incorporating perturbation stability leads to significantly better separation in the learned feature space.

be practical or accessible in real-world scenarios. Additionally, Normia focuses primarily on LLMs with a range of parameter sizes, but it remains to be seen how it performs across other types of models or in scenarios with extreme data sparsity. Another limitation is the computational cost associated with gradient norm calculations, which may increase for larger models or datasets. Finally, although our experiments demonstrate robust results across various settings, further research is needed to assess the generalizability of Normia to unseen types of LLMs and its ability to withstand countermeasures designed to mitigate MIA.

A.6. Ablation Studies

To further understand the mechanisms behind the success of OR-MIA, we conduct ablation studies focusing on the key components of our method.

The Impact of Perturbation. We investigate the direct impact of the perturbation strategy outlined in Section 3.2. Specifically, we first examine the effect of the number of perturbations N. Figure 5 plots the attack accuracy as N increases from 1 (equivalent to "No Perturbation") to 15. Accuracy rises sharply as the first few perturbations are introduced, confirming that $g_2, \cdots, g_N$ add significant discriminative information beyond g_1. The performance tends to saturate around $N = 10$, suggesting that this number of perturbations adequately captures the relevant robustness profile for distinguishing members from non-members, validating our choice of $N = 10$ in the main experiments. We then examine the T-SNE of feature representations of our OR-MIA w/ vs. w/o perturbation for members and non-members on Arxiv-6B in Figure 6 . The results show that the feature space becomes more separable when perturbation information is included, which strongly supports our hypothesis in Section 3.2 that input robustness, measured via gradient norm stability under perturbation, is a key differentiator between members and non-members, and effectively harnessing this signal is critical to the high performance of OR-MIA.

The Effect of Layers. We investigate the role of model layers in membership inference attacks. Figure 4 demonstrates that deeper layers consistently yield higher attack success rates. However, an interesting phenomenon is observed: the attack accuracy peaks in the middle layers of the model. For instance, on Arxiv-3B, the accuracy increases from 0.68 in the 5th layer to 0.93 in the 15th layer, but declines slightly to 0.87 in the 30th layer. This trend can be attributed to the middle layers capturing the richest representations during training, which facilitates better differentiation between member and non-member samples. The observed phenomenon that middle layers exhibit higher attack accuracy can be attributed to the hierarchical nature of feature learning in large language models. Prior studies, such as on Transformer-based architectures, have demonstrated that middle layers capture the richest and most abstract representations (Raganato & Tiedemann, 2018; Skean et al., 2024). These representations provide better separability between member and non-member samples, which explains the peak accuracy of 0.93 at the 15th layer in Arxiv-3B, compared to 0.68 in the 5th layer and 0.87 in the 30th layer. This finding underscores the importance of targeting middle layers to maximize attack performance, as they encode more specific information that is critical for distinguishing membership status.

A.7. Justification for Fine-Tuning Membership Setup

In our experimental setup, we define member samples as those included in a controlled fine-tuning dataset. This design choice arises from a practical constraint: the pretraining data of open-source LLMs is typically massive, weakly documented, and inaccessible in labeled form. Therefore, it is infeasible to obtain reliable membership ground truth from the pretraining stage.

Table 2. Runtime Cost Analysis of OR-MIA Across Model Sizes

Model Size	Gradient Norm Calculation (s)	SVM Training (s)	Prediction (s)	Total Runtime (s)
70M	115.23	1.01	0.67	116.91
160M	143.84	1.56	0.98	146.38
1B	178.56	3.12	1.42	183.10
3B	205.47	7.09	3.11	215.67
6B	263.23	14.52	5.35	283.10

To enable rigorous evaluation of membership inference attacks, we simulate a realistic training scenario by fine-tuning the target model on a known dataset and labeling those samples as members. Although these samples originate from a fine-tuning stage, they fulfill the same functional role in training as pretraining data in standard LLM pipelines — they directly influence model parameters.

As a result, our OR-MIA attack does not merely probe fine-tuning memorization; it targets the true training set of the model under evaluation. The inference mechanism, based on optimization and robustness signals, applies equally to pretraining and fine-tuning members, provided that gradient access is available and training exposure occurred. This makes OR-MIA broadly applicable as a diagnostic tool for detecting data leakage in both pretraining and fine-tuning regimes.

A.8. Discussion: Potential Defenses

Although this work centers on exposing a novel vulnerability in LLMs via optimization and robustness-informed membership inference, it is crucial to understand how existing or prospective defenses might mitigate the threat.

Differential Privacy (DP). DP training algorithms, by design, introduce noise to gradients during model updates. This noise could obfuscate the optimization signals (i.e., gradient norm magnitudes) that OR-MIA depends on, thus potentially reducing attack effectiveness. However, strong DP guarantees often come at the cost of model utility.

Gradient Clipping and Pruning. Techniques such as gradient clipping or low-magnitude gradient pruning, widely used in large-scale LLM training for stability, may inadvertently disrupt the gradient stability signals leveraged in OR-MIA. These techniques could reduce the separability between member and non-member features in our attack's embedding space.

While we do not evaluate these defenses empirically in this work, they represent promising directions for mitigating OR-MIA-style attacks. A comprehensive benchmark of defense effectiveness remains an important avenue for future research.

A.9. Runtime Cost Analysis

While the effectiveness of OR-MIA has been clearly demonstrated, it is equally important to assess the computational cost of the attack, especially as model size increases. This section provides a quantitative analysis of the attack's runtime across various model sizes to better understand the trade-offs between attack success and computational resources.

We measure the runtime of the attack on models with parameter sizes ranging from 70M to 6B parameters. For each model, we report the time taken to compute gradient norms across a set of perturbed inputs, as well as the overall time required for training the SVM classifier and making membership predictions.

Experimental Setup. We evaluate the runtime of OR-MIA on the following model sizes: 70M, 160M, 1B, 3B, and 6B parameters. For each model, we calculate the time taken to:

- Compute the gradient norms for each perturbed input (using 10 perturbations per sample).
- Train the SVM classifier on a calibration set of 1000 samples.
- Make membership predictions for a test set of 5000 samples.

All experiments were conducted on a machine with an NVIDIA A100 GPU and an Intel Xeon CPU.

Results. The following table summarizes the average runtime (in seconds) for each of the components of OR-MIA across different model sizes.

As shown in Table 2, the time required for gradient norm calculation scales approximately linearly with the model size, as expected, since the number of parameters influences the complexity of gradient computation. Similarly, the time required for training the SVM classifier and making predictions also increases with model size, although the classifier training time grows more rapidly for larger models due to the increased number of feature vectors.

Discussion. From the results in Table 2, it is clear that OR-MIA remains computationally feasible for models up to 1B parameters. However, as model size increases beyond 3B parameters, the runtime cost grows substantially. For practical deployment, the attack could be computationally expensive for very large models, such as 6B parameters, especially if multiple attack iterations or a large number of perturbations are needed.