Robust Membership Inference for Large Language Models under Adversarial Generative Corruption

Yuanhong Huang[1*] **Huili Wang**[2*] **Xueying Bai**[1] **Jinrui Wang**[1]
Jiajun Liu[1] **Ziqin Wang**[1] **Wanchun Ni**[3] **Shangguang Wang**[1] **Tao Qi**[1†]
[1]Beijing University of Posts and Telecommunications
[2]Tsinghua University [3]ETH Zurich

Abstract

Membership inference attack (MIA) has emerged as a promising tool for auditing the training data of LLMs, supporting data privacy and copyright protection. Most existing MIA methods rely on the assumption that LLMs assign higher confidence scores to training samples than to non-training ones. However, since LLMs generate text by sampling high-confidence tokens, they naturally produce AI-generated texts (AIGTs) that also satisfy this assumption. In this work, we empirically confirm that such AIGTs, regardless of whether they are generated by the target LLM, can lead existing MIAs to assign even higher membership likelihoods than those of true training samples, thereby significantly undermining their reliability. To address this challenge, we propose a robust membership inference framework for reliably identifying training data. Our method adopts a mixture-of-experts formulation to jointly model interactions across complementary features derived from multiple MIA methods and AIGT detectors, which can remain robust against adversarially generated samples. Furthermore, by leveraging expert components, our method provides explainable insights into the characteristics of member data. Experiments on various datasets and LLMs show that adversarial samples substantially degrade the performance of baselines, whereas our method preserves performance close to that of the unattacked setting. Codes and datasets are released at https://github.com/kong-hyh/MoMIA.

1 Introduction

With the rapid advancement of LLMs, safeguarding the privacy and intellectual property of training data has become a critical challenge, as these models are typically trained on massive and opaque corpora that hinder effective auditing of unauthorized data usage. Membership inference attacks offer a principled mechanism for auditing such risks by assessing whether specific data samples were involved in model training (Hu et al., 2022). Existing studies have extensively explored MIAs across diverse model paradigms, from discriminative classifiers to generative models, and largely converge on a shared principle: leveraging discrepancies in model confidence between training and non-training data (Shokri et al., 2017; Carlini et al., 2022). Accordingly, recent MIAs targeting LLMs inherit this paradigm and primarily rely on confidence-related signals for detection, such as logits and token-level perplexity (Shi et al., 2023).

In this paper, we demonstrate that although confidence-based principles have proven effective for MIA on traditional classification models, they exhibit a fundamental vulnerability in the context of LLMs. Specifically, LLMs can generate fluent textual content through high-confidence token sampling mechanisms (Sivaprasad et al., 2025), producing outputs that closely resemble training samples under the assumptions exploited by existing MIAs. Such AI-generated texts therefore act as a strong confounding factor, fundamentally undermining the reliability of confidence-based MIA techniques. Our empirical results in Fig. 1 substantiate this observation: when existing MIA methods (Shi et al., 2023; Zhang et al., 2025; Wang et al., 2025) are tasked with distinguishing true training samples from model generated content, they frequently assign higher membership likelihood to the generated samples, even when these samples are produced by other models, while failing to correctly identify the ground-truth training data. This limitation is particularly problematic when MIAs are deployed as forensic tools for auditing data usage and potential leakage.

To address this challenge, a straightforward strategy is to first apply existing AIGT detection

*Equal contribution.
†Corresponding Author. (Email: taoqi.qt@gmail.com)

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39531–39547

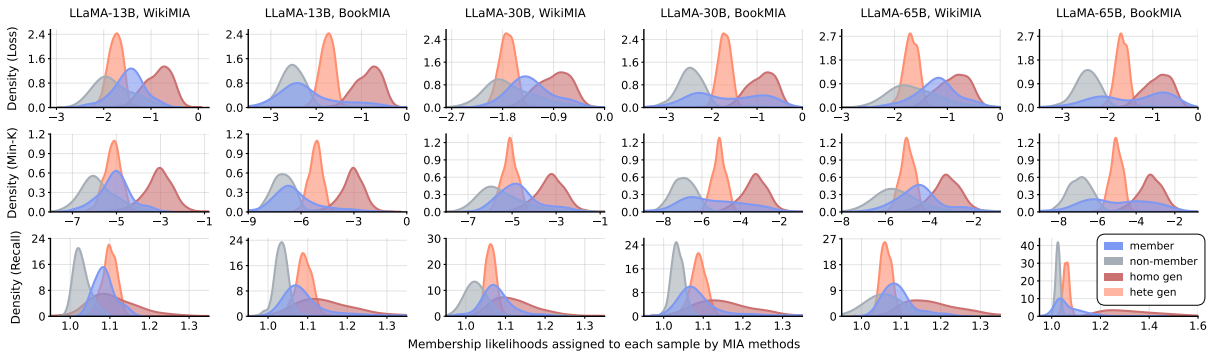

Figure 1: The distribution of membership likelihood scores for natural member samples, natural non-member samples, and adversarial samples generated by both the target (*homo gen*) and non-target LLMs (*hete gen*). The results show that all three representative MIA baselines consistently assign higher membership likelihoods to generated samples than to genuine member samples, confirming that AIGTs act as effective adversarial inputs.

methods (Mireshghallah et al., 2022; Yang et al., 2024) to filter out adversarial samples, and then perform membership inference on the remaining data. However, this two-stage approach is limited by the non-ideal performance of current AIGT detectors. Specifically, overly sensitive detectors tend to mistakenly discard genuine member samples, resulting in substantial false negatives of training data detection. Conversely, detectors with lower sensitivity allow a substantial number of adversarially generated texts to bypass filtering, thereby contaminating the subsequent membership inference stage. Thus, this unavoidable sensitivity–specificity trade-off in the two-stage framework leads to sub-optimal robustness of membership inference under adversarial perturbations.

In this paper, we propose a robust MIA framework MoMIA that enhances the reliability of membership inference in the presence of adversarially generated samples. Our key insight is to explicitly explore and enhance the distinguishability between true member samples and adversarial generations, rather than attempting to improve the overall performance of AIGT detection. We treat each linear combination of a membership inference algorithm and an AIGT detection method as an expert, and nonlinearly aggregate multiple such experts within a mixture-of-experts formulation. By jointly modeling heterogeneous MIA and AIGT signals, the proposed framework is able to capture subtle and high-order feature interactions that distinguish true member samples from adversarially generated ones. Moreover, the decoupled structure of individual experts provides interpretable insights into how different evidential cues contribute to the final decision, enabling explainable and fine-grained forensic analysis for detecting unauthorized use of training data. We conduct experiments on various benchmarks and LLMs, under adversarial samples generated by both homogeneous and heterogeneous LLMs. Results show that adversarial samples can severely degrade SOTA MIA methods to near-random guessing, whereas our method maintains performance comparable to the unattacked setting, demonstrating strong robustness against adversarial interference. Our main contributions are as follows:

(1) We are the first to demonstrate that generated texts act as strong confounders, causing SOTA MIAs to misclassify AIGTs as training members, which fundamentally undermines their reliability.

(2) We propose a robust MIA framework to provide interpretable insights for distinguishing training members from adversarial samples.

(3) Experiments show that while adversarial AIGT samples degrade existing MIAs to near-random guessing, our framework maintains performance comparable to the unattacked setting.

2 Rethink MI Attack for LLMs

2.1 Primary Experimental Setups

Models & Datasets. Most LLMs do not disclose details on training corpus, which complicates reliable MI evaluation. An exception is the LLaMA-1 family, whose training corpus has been officially documented and is widely adopted in prior studies. Following them (Shi et al., 2023), we employ three LLaMA-1 models (i.e., 13B, 30B, and 65B) and corresponding benchmarks (WikiMIA and BookMIA), for evaluation. Furthermore, additional models and their corresponding experimental results are provided in the Appendix A.7.

Adversarial Sample. We consider two adversarial generation settings. (1) Homogeneous setting (*homo gen*): adversarial samples are generated by the target LLM, representing a worst-case scenario where attackers exploit the audited model. (2) Heterogeneous setting (*hete gen*): samples are produced by an external model (GPT-4o), capturing a more realistic cross-model attack. To align distributions, we summarize each member sample into one sentence and regenerate a text from it. This preserves semantic content while removing verbatim overlap, enabling fair evaluation.

MIA Baselines. (1) Min-K% (Shi et al., 2023): aggregate the lowest token probabilities to capture memorization. (2) Min-K%++ (Zhang et al., 2025): improve Min-K% via refined token-level normalization. (3) ReCall (Xie et al., 2024): a contrastive retrieval-based method exploiting recall consistency under input perturbations. (4) Con-Recall (Wang et al., 2025): enhance ReCall via contextual consistency modeling. All these methods rely on the assumption that LLMs assign higher confidence to training samples.

2.2 Detection Logits Distribution Illustration

We present the MIA likelihood distributions for four sample categories in Fig. 1, including true members, non-members, as well as homogeneous and heterogeneous AIGTs. Distributions results of more methods are provided in Appendix A.3. First, training samples receive higher membership confidence than non-members, confirming MIA effectiveness under benign settings. Second, we observe that AIGTs are usually assigned even higher membership confidence than ground-truth training samples. This phenomenon arises because AIGTs are produced via iterative sampling of high-confidence tokens, which aligns closely with the confidence-based assumptions exploited by existing MIAs. As a result, LLMs naturally exhibit elevated confidence on AIGTs, leading to systematic misclassification and substantially undermining the reliability of MIA methods. Third, we find that this issue persists even when the AIGTs are produced by a non-target model. In such cases, MIA methods still tend to assign higher membership confidence to generated samples than to true training data, demonstrating that this vulnerability generalizes across both homogeneous and heterogeneous generations. Overall, these findings demonstrate that AIGTs can severely undermine the reliability of MIAs.

3 Methodology

3.1 Motivation

A naive defense strategy applies an off-the-shelf AIGT detector to filter adversarial samples, followed by membership inference on the remaining data. However, the non-ideal accuracy of existing AIGT detectors inevitably entangles natural samples with AI-generated ones (Fig. 2 A and Appendix A.4). This induces an inherent trade-off: a high detection threshold fails to remove all adversarial samples, whereas a low threshold discards genuine members and leads to severe false negatives. Importantly, our goal is not to separate AIGTs from natural data, but to accurately identify member samples from a mixed set containing generated samples and natural non-members.

As illustrated in Fig. 1 and Fig. 2 (A), MIA methods provide discriminative signals for separating members from AIGT samples, while AIGT detectors offer complementary cues for distinguishing non-members. This complementarity stems from their shared reliance on model confidence, yet at different granularities. MIA methods exploit fine-grained, token-level confidence statistics, whereas AIGT detectors primarily rely on coarse, sample-level confidence measures. These observations suggest that modeling interactions across heterogeneous confidence features can expose more intrinsic characteristics of member samples, enabling robust inference. Although over-parameterized neural networks are capable of modeling such interactions, their lack of interpretability undermines the trustworthy use of MIA as forensic evidence for auditing training data misuse and, moreover, limits the insights needed to inform and inspire further research. To address this issue, we first construct interpretable LightGBM that capture transparent combinations of MIA and AIGT features. We then integrate multiple such experts through a non-linear routing mechanism to model higher-order interactions. The decoupled expert structure preserves interpretability while maintaining expressive power, yielding principled and explainable membership inference.

3.2 Framework Overview

Next, we briefly introduce the mixture-of-experts–based framework for robust membership inference attacks, termed MoMIA. Given a suspected sample x, MoMIA first applies existing MIA methods and AIGT detection methods to extract corre-

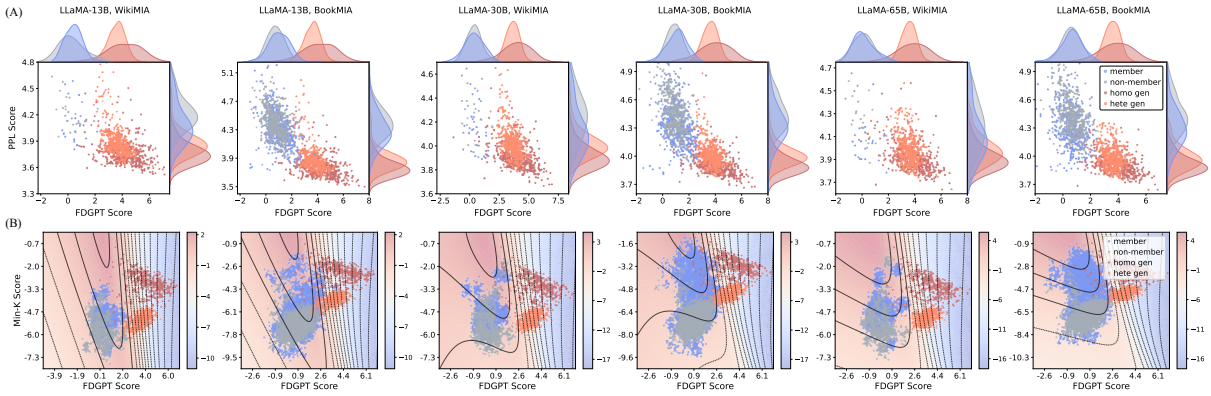

Figure 2: (A) Distribution of AIGT features from PPL (Hashimoto et al., 2019) and FDGPT (Bao et al., 2024) across member, non-member, homogeneous, and heterogeneous samples. Significant overlap is observed between member and adversarial samples, indicating limited separability. (B) Data distribution in the joint space of an MIA feature (Min-K%) and an AIGT feature (FDGPT), together with the decision boundaries learned by MoMIA decomposed into three expert models, showing that feature interactions provide insights for identifying members.

sponding feature representations, denoted as $\mathbf{m}_x$ and $\mathbf{d}_x$, respectively. Both of the feature vectors can be produced by arbitrary off-the-shelf MIA or AIGT detection algorithms. To model interactions between these features and enable more robust membership inference, MoMIA employs a set of linear expert models $\mathcal{E}_i(\mathbf{m}_x, \mathbf{d}_x; \mathbf{w}_i)$, each parameterized by $\mathbf{w}_i$. These experts are combined through a learned expert routing mechanism $\boldsymbol{\pi}(\cdot)$, yielding the final prediction:

$$\hat{y} = \sigma\left\{\sum_{i=1}^{K} \boldsymbol{\pi}_i \cdot \mathcal{E}_i(\mathbf{m}_x, \mathbf{d}_x)\right\},\ 1 \le i \le K, \quad (1)$$

where $\hat{y}$ is the output logits for membership inference, $\boldsymbol{\pi}_i$ is the routing weight for the i-th expert, and $\sigma(\cdot)$ is the sigmoid function. The routing weights are computed by a routing model:

$$\boldsymbol{\pi}(\mathbf{m}_x, \mathbf{d}_x) = \text{softmax}(\mathbf{U}^m\mathbf{m}_x + \mathbf{U}^d\mathbf{d}_x + b), \quad (2)$$

where $\{\mathbf{U}^m, \mathbf{U}^d, b\}$ are learnable parameters of the expert router. In this way, MoMIA can capture underlying interactions between these heterogeneous features to represent the characteristics of member samples, enabling robust membership inference under adversarially generated samples.

3.3 Insights from Decoupled Experts

Importantly, the mixture-of-experts architecture is inherently decomposable. The routing weights π explicitly quantify the contribution of each expert to the final decision, enabling the identification and pruning of experts with consistently low importance. By discarding such low-weight experts, MoMIA not only preserves robustness against adversarially generated samples, but also provides interpretable and fine-grained insights into how heterogeneous feature interactions support membership inference. Specifically, Fig. 2 (B) visualizes the data distribution in the joint space of an MIA feature (Min-K%) and an AIGT feature (FDGPT), together with the decision boundaries learned by MoMIA decomposed into three expert models, from which we derive three insights.

Insights. First, samples assigned extremely high AIGT scores are highly likely to be adversarially generated and should therefore be filtered out. Second, samples with moderate AIGT confidence are inherently ambiguous, as they may correspond to either generated data or genuine training members. This ambiguity arises because AIGT scores, like MIA scores, are derived from sample-level model confidence. Importantly, within this regime, generated samples consistently attain higher MIA confidence, indicating that they better satisfy the core assumptions of existing MIA methods. Consequently, samples exhibiting moderate, rather than high, MIA confidence in this region should be identified as members. Third, when AIGT scores are low, most adversarially generated samples have already been excluded. In this case, samples with high MIA confidence can be reliably recognized as members.

Discussion. The observations suggest that MIA under adversarial generation is fundamentally an inference problem under mixed data distributions, where natural non-members, true training samples, and generated texts coexist. From this per-

Model	Method		BookMIA					WikiMIA					AUC_{avg}
			Norm	HoMix	HeMix	SemiHo	SemiHe	Norm	HoMix	HeMix	SemiHo	SemiHe	
LLaMA-13B	Naive MIA	Min-K%	67.00	3.47	18.22	35.68	43.09	76.39	1.39	53.47	43.75	67.36	$40.98_{(-50.67)}$
		Min-K%++	65.26	2.21	14.97	34.38	40.36	83.33	8.33	25.00	41.67	50.00	$36.55_{(-55.10)}$
		ReCall	92.43	27.31	42.29	58.89	67.70	97.92	30.56	31.25	70.14	65.28	$58.38_{(-33.27)}$
		Con-ReCall	90.20	27.22	45.25	57.96	68.05	97.92	75.69	61.11	88.19	78.47	$69.01_{(-22.64)}$
	Two-Stage MIA	Min-K%	67.00	72.92	65.66	68.68	67.60	76.39	76.39	90.28	75.00	89.58	$74.95_{(-16.70)}$
		Min-K%++	65.26	73.09	68.30	67.65	67.29	83.33	75.00	91.67	75.00	83.33	$74.99_{(-16.65)}$
		ReCall	92.43	62.43	58.11	75.27	76.19	97.92	75.69	83.33	55.56	63.19	$74.01_{(-17.64)}$
		Con-ReCall	90.20	58.93	56.46	72.34	74.41	97.92	88.19	85.42	77.08	80.56	$78.15_{(-13.50)}$
	LightGBM		64.72	82.95	87.24	74.41	76.83	91.32	96.53	95.49	94.44	93.75	$85.77_{(-5.88)}$
	MoMIA		79.70	92.53	99.06	84.14	89.51	91.67	93.75	100.00	93.06	93.06	$91.65_{(0.00)}$
LLaMA-30B	Naive MIA	Min-K%	81.10	13.62	43.70	47.15	62.89	87.76	9.18	60.71	54.08	75.51	$53.57_{(-38.25)}$
		Min-K%++	75.01	5.06	27.92	41.09	51.93	87.76	0.51	28.06	46.94	56.63	$42.09_{(-49.73)}$
		ReCall	90.10	24.53	43.18	55.28	66.13	88.78	6.12	55.61	51.53	71.43	$55.27_{(-36.56)}$
		Con-ReCall	94.49	25.77	58.81	58.05	76.27	94.90	27.04	80.10	65.31	86.73	$66.75_{(-25.08)}$
	Two-Stage MIA	Min-K%	81.10	68.30	78.60	72.90	79.72	87.76	78.57	90.82	60.20	82.14	$78.01_{(-13.81)}$
		Min-K%++	75.01	67.36	77.42	69.96	75.61	87.76	75.00	83.16	57.14	68.37	$73.68_{(-18.15)}$
		ReCall	90.10	62.20	65.77	72.64	77.50	88.78	68.37	87.76	44.39	76.02	$73.35_{(-18.47)}$
		Con-ReCall	94.49	62.67	75.89	75.02	84.70	94.90	77.04	95.41	56.63	90.31	$80.71_{(-11.12)}$
	LightGBM		87.98	77.69	82.12	82.28	84.97	85.20	68.88	93.11	77.55	87.24	$82.70_{(-9.12)}$
	MoMIA		84.14	87.98	99.85	83.88	92.00	89.80	88.27	100.00	93.88	98.47	$91.83_{(0.00)}$
LLaMA-65B	Naive MIA	Min-K%	85.76	20.13	50.39	52.28	68.24	81.94	14.58	62.50	54.17	75.69	$56.57_{(-34.09)}$
		Min-K%++	79.22	7.44	29.98	44.13	54.17	79.86	2.08	15.28	40.28	45.83	$39.83_{(-50.83)}$
		ReCall	78.01	1.60	42.75	40.50	60.50	86.81	6.94	71.53	53.47	84.72	$52.68_{(-37.98)}$
		Con-ReCall	68.32	0.78	42.61	35.05	55.06	93.75	13.19	82.64	59.03	90.28	$54.07_{(-36.59)}$
	Two-Stage MIA	Min-K%	85.76	59.01	66.80	70.15	76.66	81.94	74.31	90.97	41.67	84.72	$73.20_{(-17.46)}$
		Min-K%++	79.22	54.99	55.78	65.06	67.22	79.86	64.58	76.39	34.72	61.11	$63.89_{(-26.77)}$
		ReCall	78.01	51.39	61.57	62.83	70.15	86.81	74.31	98.61	35.42	88.89	$70.80_{(-19.86)}$
		Con-ReCall	68.32	51.38	61.95	57.77	64.98	93.75	77.08	100.00	38.89	93.06	$70.72_{(-19.94)}$
	LightGBM		66.44	81.81	91.72	75.01	80.27	55.56	84.03	99.65	76.39	82.99	$79.39_{(-11.27)}$
	MoMIA		81.16	93.62	99.41	86.78	90.77	79.86	92.36	100.00	88.89	93.75	$90.66_{(0.00)}$

Table 1: Membership inference performance across two benchmarks, three LLMs, and four adversarial settings. Results show that our method consistently enhances MIA robustness.

spective, effective auditing requires estimating the membership posterior conditioned on heterogeneous evidential cues, rather than making sequential or hard decisions based on any single signal alone. A principled formulation is to directly model the probability of membership conditioned on both MIA- and AIGT-related features, allowing generation likelihood to serve as a contextual variable that modulates, rather than overrides, membership evidence. This unified probabilistic perspective highlights the importance of modeling nonlinear interactions and distribution-dependent decision boundaries, and points toward future methods explicitly designed to remain reliable under adversarial attacks.

4 Experiment

4.1 Experiment Setups

Data Distribution. In practice, the proportions of adversarial samples and the models used to generate them are typically unknown. To simulate realistic evaluation settings, we construct test sets as mixtures of four sample types: natural member samples, natural non-member samples, homogeneous model-generated samples, and heterogeneous model-generated samples. Let $\mathbf{r}$ denote the mixture proportions of these four categories, respectively. We evaluate five representative configurations to assess robustness under different degrees of adversarial contamination. (1) *Norm* ($\mathbf{r} = 1{:}1{:}0{:}0$): the conventional MIA setting without generated samples. (2) *HoMix* ($\mathbf{r} = 1{:}0{:}1{:}0$): negative samples are fully replaced by homogeneous generations. (3) *HeMix* ($\mathbf{r} = 1{:}0{:}0{:}1$): negatives are fully replaced by heterogeneous generations. (4) *SemiHo* ($\mathbf{r} = 1{:}0.5{:}0.5{:}0$): half of the negatives are replaced by homogeneous generations. (5) *SemiHe* ($\mathbf{r} = 1{:}0.5{:}0{:}0.5$): half of the negatives are replaced by heterogeneous generations.

Baseline MIA Methods. We evaluate three cat-

Feature	Method	LLaMA-13B			LLaMA-30B			LLaMA-65B		
		BookMIA	WikiMIA	ΔAUC	BookMIA	WikiMIA	ΔAUC	BookMIA	WikiMIA	ΔAUC
Min-K% & FDGPT	LightGBM	70.81	65.49	+16.68	75.52	78.67	+13.99	76.37	85.35	+7.05
	MoMIA	82.01	87.64		88.81	93.37		86.38	89.44	
Min-K% & DNA-GPT	LightGBM	69.38	60.49	+8.79	54.33	77.86	+11.48	61.85	62.43	+4.78
	MoMIA	69.39	78.06		73.52	81.63		66.47	67.36	
Min-K% & Perplexity$_4$	LightGBM	68.42	53.96	+1.49	66.27	52.30	+18.29	70.69	63.19	+0.03
	MoMIA	67.58	57.78		73.52	81.63		66.58	67.36	
ReCall & FDGPT	LightGBM	75.98	87.92	+8.11	79.36	87.35	+2.06	74.62	76.53	+5.97
	MoMIA	84.98	95.14		85.31	85.51		79.61	83.47	
ReCall & DNA-GPT	LightGBM	59.38	71.88	+12.53	64.98	68.32	+10.01	59.61	65.56	-2.60
	MoMIA	74.65	81.67		73.01	80.31		65.93	54.03	
ReCall & Perplexity$_4$	LightGBM	61.00	65.14	+9.24	70.08	51.53	+2.21	70.20	60.62	-3.85
	MoMIA	75.30	69.31		74.31	51.73		67.85	55.28	

Table 2: Performance of MoMIA under varying feature combinations, demonstrating its generalization capability.

egories of baseline MIA methods in our experiments. (1) Naive MIA directly applies existing membership inference attacks without any additional calibration, including Min-K% (Shi et al., 2023), Min-K%++ (Zhang et al., 2025), ReCall (Xie et al., 2024), and Con-Recall (Wang et al., 2025). (2) Two-Stage MIA adopts SOTA AIGT detection methods to first filter adversarial samples, and subsequently applies MIA methods to identify member samples. (3) LightGBM-MIA, a strong and competitive feature-combination baseline, integrates signals from both MIA methods and AIGT detectors via an efficient gradient boosting tree model. Specifically, the MIA feature set includes Min-K%, Min-K%++, ReCall, and Con-Recall, while the AIGT-related feature set comprises FDGPT (Mitchell et al., 2023), DNA-GPT (Yang et al., 2024), and perplexity (Hashimoto et al., 2019). Additional results of more feature-combination baselines are provided in Appendix A.9.

Configuration of MoMIA. MoMIA employs the same set of input features as the baselines to ensure a fair comparison. The number of linear experts, denoted by K, is set to 3, which, as analyzed in Section 4.7, achieves an effective balance between model capacity and detection performance. During training, we use a learning rate of 1×10^{-3} and optimize the model for 400 epochs to ensure convergence. To promote stable and robust learning under adversarial contamination, the training data are constructed using a mixture ratio of $\mathbf{r} = (2{:}1{:}0.5{:}0.5)$, which controls the proportion of different data sources and encourages the model to generalize across diverse scenarios. Finally, we adopt the AUC score as the primary evaluation metric for all MIA methods.

4.2 Performance Evaluation

We conduct a comparative analysis of different methods across multiple experimental settings. From the result shown in Table 1, we have three main findings. First, existing MIA methods failed to effectively classify member samples when AI-generated content is involved. For example, compared to standard MIA tasks, existing MIA methods exhibit an average AUC drop of 67.31% on *HoMix* and 42.24% on *HeMix*. This highlights the urgent need to improve existing MIA methods to enhance their robustness when AI-generated texts are incorporated. Second, MoMIA achieves consistently high AUC across multiple datasets, indicating strong generalization capability. Compared with the baseline methods, it delivers substantial performance gains: up to 39.23% over Naive MIA, 17.51% over Two-Stage MIA, and 8.76% over the LightGBM-MIA. This indicates that MoMIA not only integrates the strengths of all MIA features but also effectively captures the nonlinear relationships between AIGT and MIA, thereby improving discriminative power and robustness. Third, MoMIA maintains a strong discriminative capability under different degrees of adversarial contamination. For instance, for *HoMix* setting in BookMIA, all Naive MIA methods yield AUC below 0.5, indicating that their predictions are essentially reversed for this data split and thus completely ineffective. In contrast, MoMIA achieves an AUC

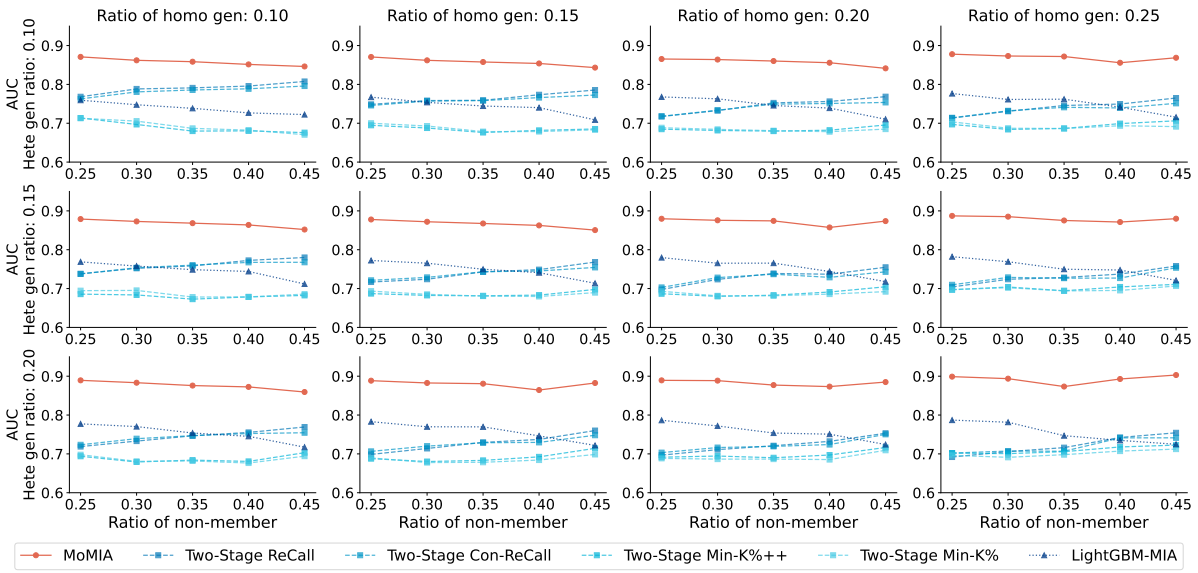

Figure 3: Performance of MoMIA under varying data mixtures, i.e., different ratios of member, non-member data, and heterogeneous and homogeneous generations, showing its robustness against diverse adversarial strategies.

of 92.53%, demonstrating exceptional discriminative capability in this challenging scenario.

4.3 Feature Generalization of MoMIA

We evaluate the discriminative capability of MoMIA and LightGBM-MIA under different combinations of AIGT features and MIA features (Table 2 and Appendix A.5). First, the results show that MoMIA outperforms the LightGBM-MIA method across almost all feature combinations. For example, MoMIA achieves an average AUC improvement of 9.47% on LLaMA-13B and 9.67% on LLaMA-30B. Second, MoMIA maintains strong and stable performance across diverse feature combinations and model scales, indicating superior generalization capability under varying feature interactions. For example, across different LLaMA model scales, MoMIA consistently achieves AUC improvements under all evaluated feature combinations. The consistent improvements observed across different feature combinations and model sizes demonstrate that MoMIA is not sensitive to specific feature choices, highlighting its robustness and general applicability in practical membership inference scenarios.

4.4 Robustness Analysis

We evaluate the performance of methods under different degrees of adversarial contamination (Fig. 3 and Appendix A.6). The results show that MoMIA consistently exhibits stable and superior performance under all degrees of adversarial contamination, significantly outperforming all baseline methods overall. For example, when the adversarial contamination changes from $\mathbf{r} = (0.3{:}0.25{:}0.25{:}0.2)$ to $\mathbf{r} = (0.05{:}0.5{:}0.25{:}0.2)$, which implies a continuous decrease in the proportion of member samples, the AUC of MoMIA remains stable without evident performance degradation or large fluctuations. These observations indicate that the proposed MoMIA method is highly robust to changes in data composition and can maintain strong discriminative capability in various complex and challenging data environments.

4.5 Generalization across Training Sources

In this section, we evaluate the cross-dataset generalization ability of MoMIA (Fig. 4). In realistic scenarios, MIAs often encounter data from unseen domains. To simulate this setting, we train MoMIA on one dataset and test it on another, and compare its performance with the strongest baseline. As expected, cross-dataset evaluation introduces a natural domain shift, leading to a slight degradation in detection performance. Specifically, compared to the upper bound obtained by training and testing on the full datasets, training on WikiMIA and testing on BookMIA results in a 7.75% decrease in AUC. Nevertheless, even under such challenging out-of-domain settings, MoMIA still substantially outperforms the strongest baseline in both evaluated scenarios. This further highlights the robustness of MoMIA under distributional shifts across datasets.

Train Model	Eval Model	BookMIA					WikiMIA					AUC_{avg}
		Norm	HoMix	HeMix	SemiHo	SemiHe	Norm	HoMix	HeMix	SemiHo	SemiHe	
LLaMA-13B	LLaMA-13B	79.70	92.53	99.06	84.14	89.51	91.67	93.75	100.00	93.06	93.06	91.65
	LLaMA-30B	79.86	92.67	99.88	83.73	89.18	91.33	89.80	100.00	94.90	97.96	91.93
	LLaMA-65B	72.06	73.59	99.30	69.75	85.31	68.06	75.69	100.00	76.39	84.72	80.49
LLaMA-30B	LLaMA-13B	74.74	86.22	98.89	78.53	87.40	87.50	90.97	100.00	93.75	95.14	89.31
	LLaMA-30B	84.14	87.98	99.85	83.88	92.00	89.80	88.27	100.00	93.88	98.47	91.83
	LLaMA-65B	83.68	73.90	99.29	76.47	91.89	83.33	84.03	100.00	90.28	97.22	88.01
LLaMA-65B	LLaMA-13B	52.15	91.60	99.39	70.37	76.06	74.31	92.36	100.00	90.28	90.97	83.75
	LLaMA-30B	70.76	93.28	99.89	81.16	85.88	83.16	93.88	100.00	95.41	97.96	90.14
	LLaMA-65B	81.16	93.62	99.41	86.78	90.77	79.86	92.36	100.00	88.89	93.75	90.66
OLMo	LLaMA-13B	76.44	86.36	99.02	79.51	88.57	87.50	97.92	100.00	95.14	95.14	90.56
	LLaMA-30B	87.83	87.08	99.91	84.42	93.42	94.39	93.88	100.00	96.94	97.96	93.58
	LLaMA-65B	78.14	64.63	99.30	68.94	89.34	82.64	82.64	100.00	86.81	93.75	84.62

Table 3: Cross-model performance of MoMIA, demonstrating its strong generalization ability.

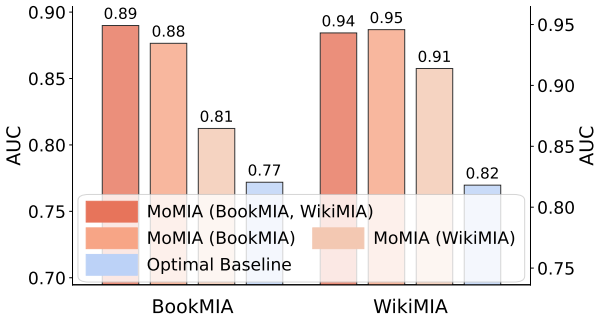

Figure 4: Generalization across training sources.

4.6 Generalization across Different Models

We investigate the cross-model generalization ability of MoMIA by training MoMIA on annotated data from a different LLM (e.g., OLMo (Groeneveld et al., 2024)) and then applying the learned feature interaction patterns to audit LLaMA-series models. As summarized in Table 3, MoMIA exhibits strong cross-model generalization performance, which suggests that it captures model-agnostic interaction structures instead of overfitting to a specific target model. These cross-model results together with the cross-dataset results, demonstrate that MoMIA works under realistic proxy supervision rather than relying on unrealistic complete label information.

4.7 Algorithm Analysis of MoMIA

In this section, we first analyze the impact of training sample size on MoMIA's performance. We vary the number of training samples and evaluate the AUC across all datasets. The results in Fig. 5 (A) show a clear trend that performance steadily improves as more training data are used. Notably, MoMIA maintains strong performance even with a relatively small number of training samples (256), achieving results close to those obtained with the full dataset. This indicates that MoMIA is data-efficient and can generalize well even under limited training supervision. Next, we further analyze the effect of the number of experts in MoMIA on its performance, as illustrated in Fig. 5 (B). We vary the number of experts from 1 to 100 and evaluate the MIA performance. The results show that, considering both performance and computational cost, setting the number of experts to 3 achieves the best trade-off.

5 Related Work

Membership Inference Attack. A large body of work on MIAs (Shokri et al., 2017; Carlini et al., 2022) has primarily been developed to quantify data leakage in traditional classification models (Mireshghallah et al., 2022; Truex et al., 2019). Their core idea is to exploit the model's stronger fit to its training data by examining confidence-based signals, such as prediction loss and logits (Ye et al., 2022; Liu et al., 2022). Inspired by this paradigm, recent studies extend confidence-based MIAs (Fu et al., 2024; Xie et al., 2024) target on LLMs. For example, Shi et al. (2023) leverage fine-grained token-level confidence and compute the average likelihood of the $k\%$ least probable tokens. However, AIGTs are produced by sampling high-confidence tokens of LLMs, which yields strong adversarial samples and significantly undermines the reliability of existing MIAs.

AIGT Detection. AIGTs have become increasingly difficult to distinguish from human-written

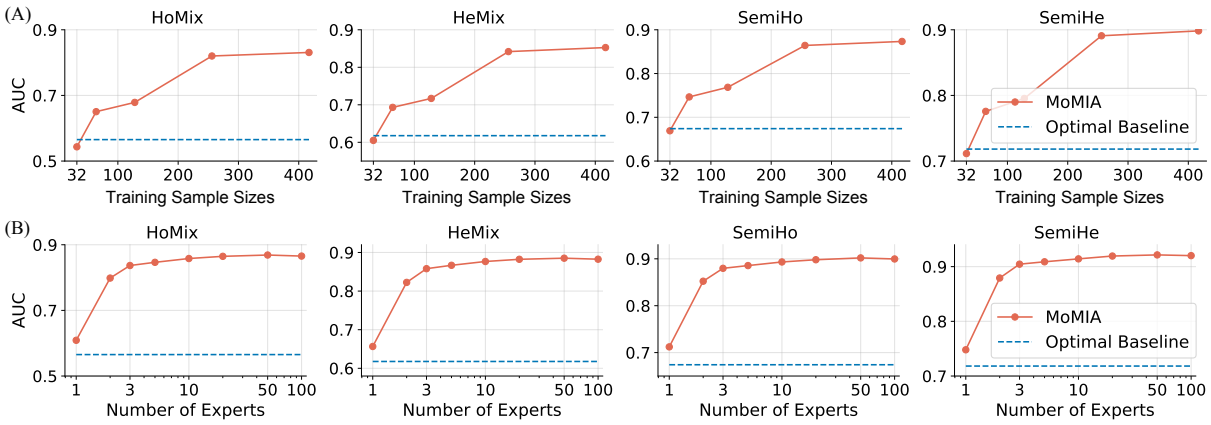

Figure 5: Algorithm Analysis. (A) Performance of MoMIA with varying training sample sizes. Results show that a small amount of data is sufficient for MoMIA to learn effective feature interaction patterns and substantially improve MI performance under adversarial settings. (B) Performance of MoMIA with varying numbers of experts.

content, prompting extensive research on AIGT detection (Park et al., 2025; Zhou et al., 2025; Mitchell et al., 2023; Bao et al., 2024; Yang et al., 2024). For example, Mitchell et al. (2023) identify AIGTs by measuring the curvature of the log-probability function under small input perturbations. While these methods are effective at separating AIGTs from natural texts, we show that their direct application fails to reliably distinguish AIGTs from true member samples (Fig. 2), yielding only limited robustness improvements for MIA under adversarial corruption. Moreover, since AIGT detectors are also built upon confidence-based signals, we show that they encode complementary cues relevant to MIA. Motivated by this observation, MoMIA jointly models signals from MIAs and AIGT detectors to explicitly capture their complex interactions, thereby enabling more reliable membership inference.

6 Conclusion

In this paper, we reveal a fundamental vulnerability of existing MIAs for LLMs: AIGTs can act as adversarial confounders, thereby undermining the core confidence-based assumption on which these methods rely. We further demonstrate that this phenomenon consistently generalizes across multiple datasets, foundation LLMs, and adversarial generation strategies, confirming a serious threat to the practical reliability of MIAs. To address this challenge, we observe that the interactions between MI signals and AIGT detection features provide discriminative evidence for identifying true member samples under adversarial corruption. Building on this insight, we propose a robust MIA framework that explicitly models such cross-feature interactions. Experiments on various datasets and LLMs demonstrate that the proposed framework achieves consistent robustness across diverse adversarial attack strategies.

Limitations

Despite the insights provided by this study, our method is not without limitation. Although extensive experiments validate the correctness of our findings and the effectiveness of the proposed method, our evaluation is limited to open-source models. We do not include more recent closed-source LLMs, primarily because their training corpora are not publicly disclosed, which constitutes a common challenge for conducting reliable evaluations of membership inference attacks on LLMs. In future work, we plan to collaborate with industrial research teams to further investigate the generalization of our approach on proprietary models with well-documented training data.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62502044, 62425203; Beijing Natural Science Foundation under Grant number L253005; CCF-SANGFOR Research Fund under Grant number 20240202; Research Initiation Project for Introduced Talents of BUPT under Grant number 2025KYQD11; and the Beijing Municipal Science & Technology Commission, the Administrative Commission of Zhongguancun Science Park under Grant number Z251100003625014.

References

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: efficient zero-shot detection of machine-generated text via conditional probability curvature. In *Proceedings of International Conference on Representation Learning*, pages 24814–24836.

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)*, pages 1897–1914. IEEE.

Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. 2025. Exploring the limitations of detecting machine-generated text. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4274–4281.

Jiachen Fu, Chun-Le Guo, and Chongyi Li. 2025. Detectanyllm: Towards generalizable and robust detection of machine-generated text across domains and models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11229–11238.

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. In *Proceedings of Advances in Neural Information Processing Systems*, volume 37, pages 134981–135010. Curran Associates, Inc.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

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):1–37.

Guanhua Huang, Yuchen Zhang, Zhe Li, Yongjian You, Mingze Wang, and Zhouwang Yang. 2024. Are ai-generated text detectors robust to adversarial perturbations? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6005–6024.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. Mage: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53.

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*, pages 2085–2098.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of International Conference on Machine Learning*, pages 24950–24962. PMLR.

Hyeonchu Park, Byungjun Kim, and Bugeun Kim. 2025. Dart: An aigt detector using amr of rephrased text. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 710–721.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *Preprint*, arXiv:2310.16789.

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)*, pages 3–18. IEEE.

Sarath Sivaprasad, Pramod Kaushik, Sahar Abdelnabi, and Mario Fritz. 2025. A theory of response sampling in llms: Part descriptive and part prescriptive. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30091–30135.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar,

and 1 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788.

Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2019. Demystifying membership inference attacks in machine learning as a service. *IEEE transactions on services computing*, 14(6):2073–2089.

Cheng Wang, Yiwei Wang, Bryan Hooi, Yujun Cai, Nanyun Peng, and Kai-Wei Chang. 2025. ConReCall: Detecting pre-training data in LLMs via contrastive decoding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1013–1026, Abu Dhabi, UAE. Association for Computational Linguistics.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. 2024. ReCaLL: Membership inference via relative conditional log-likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8671–8689, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xianjun Yang, Wei Cheng, Yue Wu, Linda Petzold, William Wang, and Haifeng Chen. 2024. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. In *Proceedings of International Conference on Representation Learning*, volume 2024, pages 48572–48597.

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*, pages 3093–3106.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-k%++: Improved baseline for pre-training data detection from large language models. In *Proceedings of The Thirteenth International Conference on Learning Representations*.

Baohang Zhou, Zezhong Wang, Lingzhi Wang, Hongru Wang, Ying Zhang, Kehui Song, Xuhui Sui, and Kam-Fai Wong. 2024. Dpdllm: a black-box framework for detecting pre-training data from large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 644–653.

Yinghan Zhou, Juan Wen, Wanli Peng, Xue Yiming, Ziwei Zhang, and Wu Zhengxian. 2025. Kill two birds with one stone: generalized and robust ai-generated text detection via dynamic perturbations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8864–8875.

A Appendix

A.1 Experimental Setup

All experiments were conducted on a high-performance computing server running Ubuntu 22.04.5 LTS. The system is equipped with 128 Intel Xeon Gold 6342 CPUs (2.80 GHz) and two NVIDIA A800 PCIe GPUs with 80 GB memory each. All methods were implemented in Python 3.10 using PyTorch 2.7 as the primary deep learning framework.

A.2 AIGT Detection

In this section, we provide a detailed description of all AIGT detection methods used in this work.

FDGPT (Bao et al., 2024) is a zero-shot detection method based on conditional log-likelihood statistics. Given an input text x, its detection score is defined as:

$$\mathrm{FDGPT}(x) = \frac{\log p_\theta(x) - \tilde{\mu}}{\tilde{\sigma}}, \tag{3}$$

where $\log p_\theta(x)$ denotes the log-likelihood of x under the target language model, and $\tilde{\mu}$ and $\tilde{\sigma}$ are the estimated mean and standard deviation obtained via conditional sampling from the same model. Intuitively, FDGPT measures how atypical the input is relative to model-generated samples, with higher scores indicating a greater likelihood of being machine-generated.

DNA-GPT (Yang et al., 2024) detects AIGTs by comparing the likelihood of the observed continuation with that of alternative continuations sampled from the target model. Specifically, for a truncated prefix x', the score is computed as

$$\mathrm{DNA}(x') = \log p_\theta(y_0 \mid x') - \frac{1}{K}\sum_{k=1}^{K} \log p_\theta(y_k \mid x'), \tag{4}$$

where y_0 is the ground-truth continuation and $\{y_k\}_{k=1}^{K}$ are K continuations generated from the target model conditioned on x'. A larger score suggests that the original continuation is more consistent with human-authored text than model-generated alternatives.

Perplexity$_4$ (Hashimoto et al., 2019) is a higher-order statistic derived from the distribution of perplexity values across the dataset:

$$\mathrm{Perplexity}_4 = \frac{1}{N}\sum_{i=1}^{N} \left(\mathrm{PPL}(x_i) - \bar{\mathrm{PPL}}\right)^4, \tag{5}$$

where $\mathrm{PPL}(x_i)$ denotes the perplexity of the i-th sample, $\bar{\mathrm{PPL}}$ is the mean perplexity over the dataset, and N is the total number of samples. This metric captures the tail behavior and dispersion of perplexity values, which may help identify differences between human-written and AIGTs.

DetectAnyLLM (Fu et al., 2025) measures the discrepancy between an input text and its re-sampled variant. It constructs a more stable detection signal by re-sampling and perturbing the input text, and then measuring the discrepancy between the original text and its re-sampled variant.

A.3 Membership Scores Distribution

In this section, we provide a detailed analysis of the distributions of membership likelihood scores produced by different MIA methods. Specifically, we examine score distributions for three types of samples: natural member samples, natural non-member samples, and adversarial samples generated by both the target model (*homo gen*) and non-target models (*hete gen*). The corresponding results are presented in Fig.6. Across all evaluated methods, we observe a consistent pattern in which adversarially generated samples are assigned substantially higher membership likelihood scores than genuine member samples. This phenomenon holds for adversarial samples produced by both the target model and non-target models, indicating that the effect is not tied to a specific generator. These results further corroborate our findings in the main paper, suggesting that AIGTs systematically induce overconfident membership predictions and therefore function as effective adversarial inputs against existing MIAs.

A.4 AIGT Features Distribution

In this section, we present additional analyses of the distributions of AIGT-related features produced by different detection methods, as illustrated in Fig. 7. These results complement the main experimental findings by offering a more fine-grained view of how different detectors respond to various types of inputs at the feature level. Due to the non-ideal detection accuracy of existing AIGT detectors, the feature distributions exhibit substantial overlap between natural samples and AI-generated samples. This overlap indicates that adversarially generated samples cannot be perfectly separated from genuine data using a single detection score or threshold. This analysis further highlights the practical challenges faced

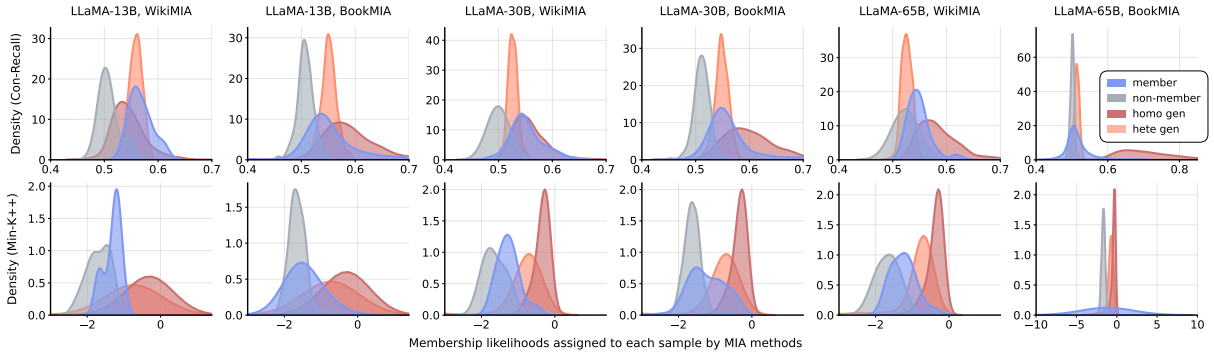

Figure 6: The distribution of membership likelihood scores of Con-Recall and Min-K++ for natural member samples, natural non-member samples, and adversarial samples generated by both the target (*homo gene*) and non-target LLMs (*hete gen*).

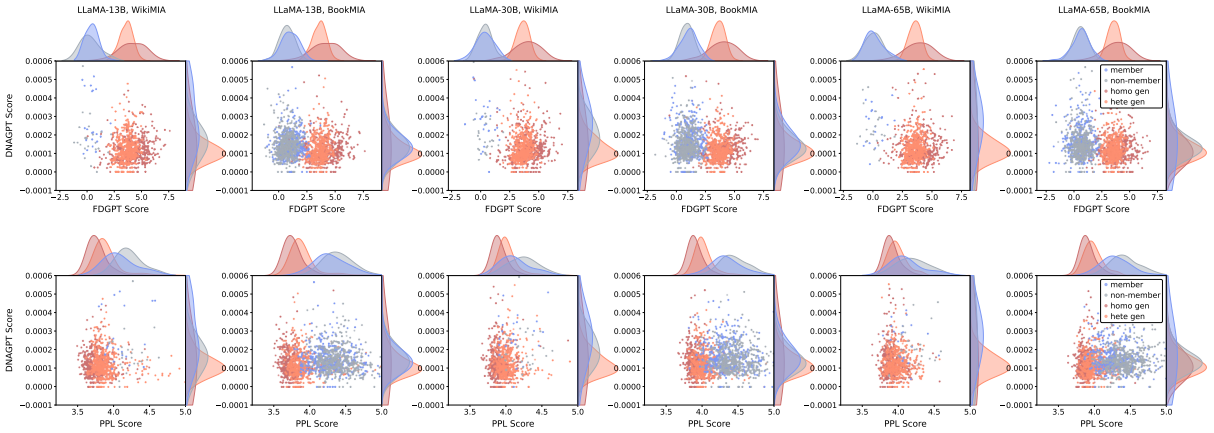

Figure 7: Distribution of AIGT features from PPL (Hashimoto et al., 2019), FDGPT (Bao et al., 2024) and DNA-GPT (Yang et al., 2024) across member, non-member, homogeneous, and heterogeneous samples. Significant overlap is observed between member and adversarial samples, indicating limited separability.

by current AIGT detection mechanisms when deployed as a pre-filtering step in MIA pipelines.

A.5 Feature Generalization of MoMIA

In this section, we present additional combinations of AIGT and MIA features to further evaluate the discriminative capability of different methods. As shown in Table 4, MoMIA consistently outperforms LightGBM-MIA across all feature combinations, demonstrating its superior discriminative power. These results complement the findings from the main experiments, indicating that MoMIA is largely insensitive to the choice of specific features, which underscores its robustness and broad applicability in practical membership inference scenarios.

A.6 Robustness of MoMIA

In this section, we further evaluate the robustness of MoMIA and baseline methods under varying degrees of adversarial contamination, as illustrated in Fig. 8. The contamination ratio controls the proportion of adversarially generated samples involved in the inference process, thereby simulating different levels of attack strength. MoMIA consistently maintains stable and superior performance across all contamination levels, significantly outperforming all baseline methods. In contrast, the performance of baseline approaches degrades noticeably as the level of adversarial contamination increases, while MoMIA exhibits only marginal fluctuations, demonstrating strong resilience to adversarial noise. These results complement the findings in the main experiments, further confirming that MoMIA is robust to adversarial interference and can reliably capture membership signals under diverse and challenging inference conditions.

A.7 Performance on Additional Models

In this section, we conduct additional experiments on OLMo-7B (Groeneveld et al., 2024) and

Feature	Method	LLaMA-13B			LLaMA-30B			LLaMA-65B		
		BookMIA	WikiMIA	ΔAUC	BookMIA	WikiMIA	ΔAUC	BookMIA	WikiMIA	ΔAUC
Min-K%++ & FDGPT	LightGBM	71.24	67.15	+12.74	76.19	82.55	+5.32	75.95	76.25	+8.69
	MoMIA	78.87	85.00		84.70	84.69		84.57	85.00	
Min-K%++ & DNA-GPT	LightGBM	69.35	63.19	+7.25	66.15	87.40	-4.57	63.02	61.94	+2.68
	MoMIA	71.07	75.97		69.60	74.80		68.52	61.81	
Min-K%++ & Perplexity$_4$	LightGBM	68.63	54.93	+4.34	75.98	69.44	-3.34	70.71	56.67	+4.03
	MoMIA	69.60	62.64		74.87	63.88		71.01	64.44	
Con-ReCall & FDGPT	LightGBM	75.11	92.36	+5.52	79.67	83.27	+10.93	72.85	74.44	+10.36
	MoMIA	88.10	90.42		90.51	94.29		78.98	89.03	
Con-ReCall & DNA-GPT	LightGBM	57.73	90.14	+7.22	67.53	78.98	+6.91	61.12	61.18	-3.23
	MoMIA	73.98	88.33		74.00	86.33		68.05	47.78	
Con-ReCall & Perplexity$_4$	LightGBM	59.02	79.44	+14.74	73.89	72.30	+0.49	71.94	60.14	-7.28
	MoMIA	79.88	88.06		75.95	71.22		69.33	48.19	

Table 4: Performance of MoMIA under varying feature combinations, demonstrating its generalization capability.

Method		Qwen3-8B					OLMo-7B					AUC$_{avg}$
		Norm	HoMix	HeMix	SemiHo	SemiHe	Norm	HoMix	HeMix	SemiHo	SemiHe	
Naive MIA	Min-K%	51.49	1.29	2.68	26.19	27.02	58.04	3.08	29.38	31.24	44.98	27.54$_{(-59.56)}$
	Min-K%++	55.41	6.09	6.10	30.80	31.49	61.65	5.77	14.01	35.35	39.31	28.60$_{(-58.50)}$
	ReCall	56.36	5.50	10.56	30.95	33.49	52.48	20.71	12.03	38.49	34.66	29.52$_{(-57.57)}$
	Con-ReCall	72.19	22.57	32.65	46.75	51.99	80.56	30.55	63.01	57.35	72.93	53.05$_{(-34.04)}$
Two-Stage MIA	Min-K%	55.67	96.95	93.90	75.83	74.18	52.65	89.57	33.47	71.15	42.20	68.56$_{(-18.54)}$
	Min-K%++	45.00	93.91	93.90	69.20	68.51	51.22	96.35	84.22	74.52	68.25	74.51$_{(-12.59)}$
	ReCall	43.68	94.50	89.44	69.05	66.51	47.52	79.29	87.97	61.51	65.34	70.48$_{(-16.62)}$
	Con-ReCall	28.61	77.43	67.35	53.25	48.01	19.44	69.45	36.99	42.65	27.07	47.02$_{(-40.07)}$
LightGBM-MIA		55.64	95.20	96.98	74.37	75.21	63.28	93.97	96.58	78.63	79.99	80.99$_{(-6.11)}$
MoMIA		65.23	98.50	98.28	80.38	80.04	76.96	97.07	99.32	86.80	88.39	87.10$_{(0.00)}$

Table 5: Membership inference performance for OLMo-7B and Qwen3-8B under four adversarial settings.

Qwen3-8B (Yang et al., 2025). For OLMo-7B, whose training corpus (Dolma (Soldaini et al., 2024)) is publicly documented, we construct the member set from Wikipedia articles included in Dolma and the non-member set from pages crawled after August 1, 2025, ensuring temporal separation from its training data. For Qwen3-8B, although its full training data is not disclosed, Wikipedia is widely recognized as part of its corpus; we therefore use pre-release Wikipedia pages as members and post-release pages as non-members, maintaining the same temporal separation. The results (Table 5) show that MoMIA achieves the highest average AUC across all mixture settings compared to baseline methods on both models. This confirms that MoMIA remains robust and effective for modern LLMs beyond LLaMA-1 (across diverse foundation models), despite differences in training pipelines or post-training alignment, demonstrating its broad applicability and practical utility.

A.8 Performance of MoMIA under varying numbers of input features

In this section, we evaluate MoMIA under varying numbers of input features with different quality levels. The results (Fig. 9) show incorporating weaker or redundant features does not significantly degrade performance. This is because the MoE router can automatically down-weight less informative features during training. Therefore, MoMIA does not depend on meticulous feature engineering; rather, it can automatically suppress low-quality features while learning meaningful feature interactions.

A.9 Comparison with Different Feature Combination Baselines

In this section, we extend our experimental comparison to include several feature combination baselines, including XGBoost, LightGBM, and

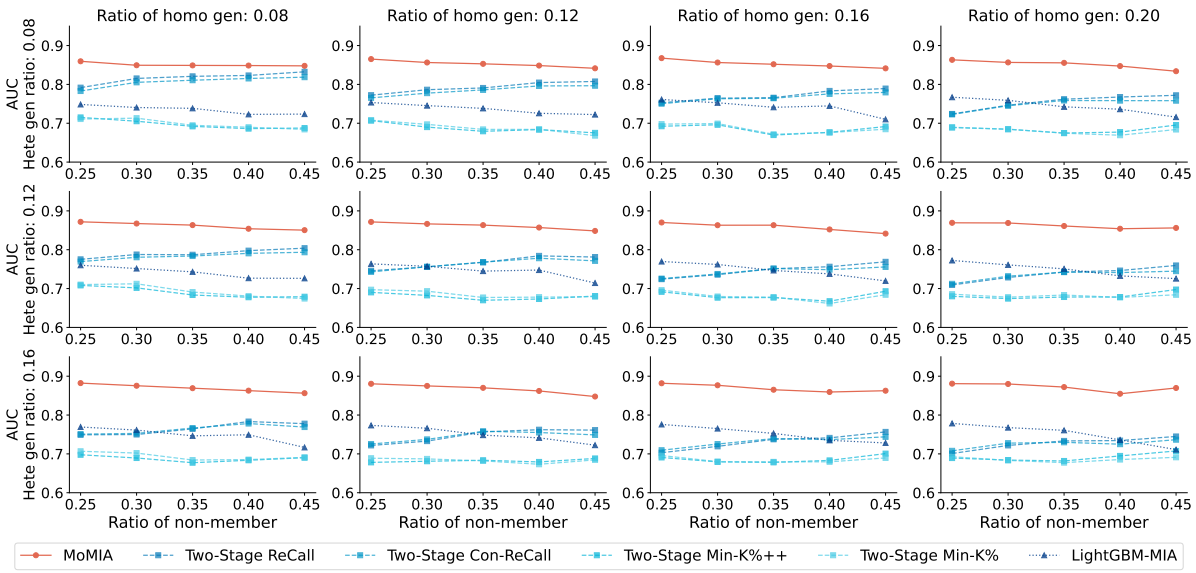

Figure 8: Performance of MoMIA under varying data mixtures, i.e., different ratios of member, non-member data, and heterogeneous and homogeneous generations, showing its robustness against diverse adversarial strategies.

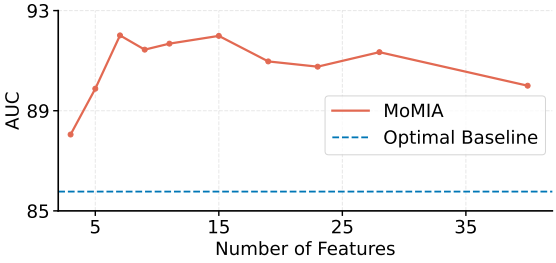

Figure 9: Results of MoMIA under varying numbers of input features with different quality levels.

MLPs, evaluated on two datasets (BookMIA and WikiMIA) and one target model (LLaMA-13B). The results reported in the Table 6 demonstrate that MoMIA consistently and significantly outperforms these alternatives. This is because the diversity of adversarial corruption patterns (e.g., variations in generative models and data mixture proportions) may introduce highly heterogeneous perturbation behaviors. Traditional statistical models (e.g., XGBoost and LightGBM) and standard neural architectures (e.g., MLPs) may struggle to capture such complex and multi-modal feature interactions. We also acknowledge that employing a sufficiently large and complex model might partially address this challenge. However, such approaches usually sacrifice interpretability and make it difficult to analyze the underlying feature interaction mechanisms driving the attack decisions. In contrast, MoMIA achieves both improved robustness and enhanced interpretability by explicitly modeling structured expert interactions. Therefore, beyond performance gains, our method contributes an explainable modeling framework for robust MIA.

A.10 Performance with Black-box MIA

To demonstrate that MoMIA is not limited to gray-box setting, we replace the current gray-box features with the SOTA black-box feature (e.g., DPDLLM (Zhou et al., 2024)). The results in Table 7 show MoMIA can be naturally extended to a purely black-box setting.

A.11 Performance of Two-Stage MIA under different detection thresholds

In this section, we conduct additional experiments to assess Two-Stage MIA performance variation under different detection thresholds. Specifically, we instantiate the two-step pipeline using FDGPT as the AIGT detector and Min-K%++ as the membership inference module evaluated on LLaMA-13B. In this framework, samples are first filtered by the AIGT detector based on a predefined detection threshold, and membership inference is then performed on the remaining samples.

We vary the FDGPT detection threshold continuously from 0 to 1. For each threshold value, we report the resulting membership inference performance measured by AUC (Table 8). The results show that the performance of Two-Stage MIA peaks at threshold = 0.5, but even under this optimal setting the AUC is substantially lower than

Method	BookMIA					WikiMIA					AUC_{avg}
	Norm	HoMix	HeMix	SemiHo	SemiHe	Norm	HoMix	HeMix	SemiHo	SemiHe	
Linear	46.14	96.05	99.60	70.61	73.54	52.78	99.31	100.00	78.47	78.47	$79.50_{(-12.15)}$
XGBoost	62.57	75.28	81.94	69.78	74.31	86.81	99.31	100.00	96.88	97.22	$84.41_{(-7.24)}$
MLP	51.85	97.86	98.72	74.59	75.79	63.89	100.00	100.00	81.94	81.94	$82.66_{(-8.99)}$
LightGBM	64.72	82.95	87.24	74.41	76.83	91.32	96.53	95.49	94.44	93.75	$85.77_{(-5.88)}$
MoMIA	79.70	92.53	99.06	84.14	89.51	91.67	93.75	100.00	93.06	93.06	$91.65_{(0.00)}$

Table 6: Comparison results of MoMIA with different feature combination baselines.

the proposed MoMIA framework. These findings demonstrate that the two-step framework is inherently unstable and heavily dependent on precise threshold tuning, which poses a practical limitation in real-world deployment. In contrast, the proposed MoMIA framework avoids such hard thresholding by jointly modeling AIGTs and MIA features within a unified architecture, resulting in consistently superior and more stable performance.

A.12 Two-Stage MIA Performance with Additional AIGT Detection Methods

In this section, we incorporate an additional state-of-the-art AIGT detection method, DetectAnyLLM (Fu et al., 2025), into the Two-Stage pipeline. The results (Table 9) indicate that although recent detectors yield marginal improvements over earlier approaches, their performance remains notably inferior to the ideal scenario and substantially worse than our proposed method. This gap arises because reliably distinguishing AI-generated texts from human-authored texts is inherently difficult. As shown in Figure 2 (A), while many samples can be correctly identified, a non-negligible fraction of long-tail generated samples overlaps with human texts, thereby degrading overall detection accuracy. Moreover, adversarial generation techniques (Huang et al., 2024; Li et al., 2024; Doughman et al., 2025) can further amplify such long-tail effects, exacerbating the performance drop of AIGT detectors. These findings highlight the necessity of modeling interaction patterns between MIA features and AIGT-related signals to achieve robust membership inference against LLMs.

Method	BookMIA					WikiMIA					AUC_{avg}
	Norm	HoMix	HeMix	SemiHo	SemiHe	Norm	HoMix	HeMix	SemiHo	SemiHe	
Naive MIA	51.18	29.31	51.09	39.49	51.84	50.69	0.00	0.69	24.31	25.00	$32.36_{(-49.71)}$
Two stage MIA	51.18	76.25	82.39	62.43	67.21	50.69	75.00	66.67	57.64	65.97	$65.54_{(-16.53)}$
LinearMIA	38.58	97.59	99.55	68.04	69.71	43.06	99.31	100.00	74.31	74.31	$76.44_{(-5.63)}$
MoMIA	51.96	96.83	99.21	73.28	75.07	65.97	97.22	100.00	80.56	80.56	$82.07_{(0.00)}$

Table 7: Black-box membership inference performance across different settings.

Method	Threshold	BookMIA					WikiMIA					AUC_{avg}
		Norm	HoMix	HeMix	SemiHo	SemiHe	Norm	HoMix	HeMix	SemiHo	SemiHe	
Two-Stage MIA	0.0	60.04	10.50	22.30	37.08	42.66	59.72	46.53	56.25	54.17	59.03	$44.83_{(-46.82)}$
	0.2	49.47	67.60	74.09	58.20	61.37	75.00	92.36	93.75	83.33	84.03	$73.92_{(-17.73)}$
	0.5	65.26	73.09	68.30	67.65	67.29	83.33	75.00	83.33	75.00	91.67	$74.99_{(-16.66)}$
	0.8	65.26	11.12	14.97	37.43	40.36	83.33	16.67	25.00	50.00	50.00	$39.41_{(-52.24)}$
	1.0	65.26	3.79	14.97	34.89	40.36	83.33	8.33	25.00	41.67	50.00	$36.76_{(-54.89)}$
MoMIA	-	79.70	92.53	99.06	84.14	89.51	91.67	93.75	100.00	93.06	93.06	$91.65_{(0.00)}$

Table 8: Membership inference performance of baseline Two-Stage MIA under different thresholds.

Method		BookMIA					WikiMIA					AUC_{avg}
		Norm	HoMix	HeMix	SemiHo	SemiHe	Norm	HoMix	HeMix	SemiHo	SemiHe	
Two-Stage MIA	Min-K%	51.03	72.28	79.01	61.78	64.94	79.86	92.36	93.75	83.33	84.03	$76.24_{(-15.28)}$
	Min-K%++	52.34	72.49	83.50	62.56	67.97	71.53	91.67	97.22	84.72	86.81	$77.08_{(-14.44)}$
	ReCall	70.92	82.69	88.35	76.56	80.02	90.28	94.44	95.83	91.67	92.36	$86.31_{(-5.21)}$
	Con-ReCall	70.18	82.07	87.66	75.99	79.19	90.28	98.61	97.22	93.06	92.36	$86.66_{(-4.86)}$
MoMIA		74.41	96.00	97.75	85.31	86.70	88.89	100.00	100.00	93.06	93.06	$91.52_{(0.00)}$

Table 9: MIA performance when DetectAnyLLM is used as the AIGT component in Two-Stage MIA. Stronger detection features limitedly improve baselines but remain below MoMIA across settings.