Enhancing Pre-training Data Detection in LLMs Through Discriminative and Symmetric Prefix Selection

Kai Sun[1,2], **Yuxin Lin**[1,2], **Bo Dong**[2,3*], **Jingyao Zhang**[4], **Bin Shi**[1,2]
[1]School of Computer Science and Technology, Xi'an Jiaotong University, China
[2]Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University, China
[3]School of Distance Education, Xi'an Jiaotong University, China
[4]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, China
sunkai@xjtu.edu.cn, yuxinlin@stu.xjtu.edu.cn, dong.bo@xjtu.edu.cn,
Jingyao.Zhang23@student.xjtlu.edu.cn, shibin@xjtu.edu.cn

Abstract

The rapid development of large language models (LLMs) has relied on access to high-quality, large-scale datasets, yet growing concerns around data privacy and security have spurred substantial research into pre-training data detection. While state-of-the-art (SOTA) methods such as RECALL and CON-RECALL leverage auxiliary prefixes to enhance detection performance, their dependence on individual prefixes introduces notable instability across varying prefix conditions. To address this, we first conduct a theoretical analysis to assess the impact of prefixes on existing prefix-based methods. Building on the analysis, we propose a novel prefix selection method to identify optimal prefixes. Specifically, our method derives two key criteria *Discriminability* and *Symmetry*. These criteria serve to quantify the effectiveness of prefixes in detecting pre-training data, enabling precise selection of high-performing candidate prefixes. Experiments on the WikiMIA dataset demonstrate that our method consistently improves the performance of RECALL and CON-RECALL, achieving gains of up to 21.1% in AUC scores while significantly enhancing robustness.

Code — https://github.com/Linyuxin03/DSC-Prefix

Introduction

In recent years, large language models (LLMs) have demonstrated remarkable capabilities across diverse language understanding tasks, including question answering (Lin et al. 2025), machine translation (Xu et al. 2024), and sentiment analysis (Miah et al. 2024). These achievements depend on massive pre-training datasets collected from diverse sources. However, as training corpora expand exponentially, the risk of models memorizing and reproducing sensitive content has emerged as a critical challenge (Meeus et al. 2024; Duarte et al. 2024; Tang et al. 2023). This concern has driven increased research interests on pre-training data detection (Xie et al. 2024; Wang et al. 2024).

The objective of pre-training data detection is to identify whether a given sample was part of an LLM's pre-training dataset. Existing approaches rely on scoring functions designed to assign a scalar value to each sample. The scores are then used to distinguish between member samples (data used during pre-training) and non-member samples (data excluded from pre-training). Therefore, the core challenge involves developing a scoring function that maximizes the score divergence between member and non-member samples.

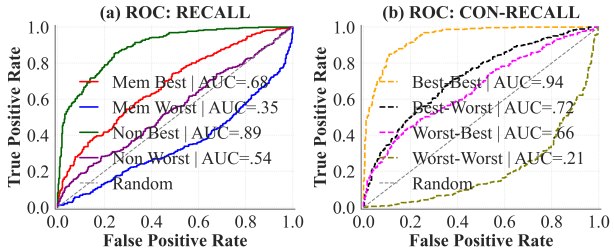

Figure 1: Performances of RECALL and CON-RECALL on WikiMIA-32 dataset across varying prefixes. **(a) RECALL:** ROC curves for best/worst-performing member prefixes and best/worst-performing non-member prefixes; **(b) CON-RECALL:** ROC curves for prefix pairs (P_0, P_1), where P_0 and P_1 are derived from RECALL's best or worst member/non-member prefixes. Four pairs (Best-Best, Best-Worst, Worst-Best, Worst-Worst) are illustrated.

Prior approaches primarily leverage probabilistic outputs from LLMs, such as log-likelihood, to construct scoring functions. For instance, Min-K% (Shi et al. 2023) calculates scores by averaging the log-likelihood of the lowest-probability tokens (k% threshold), while Min-K%++ (Zhang et al. 2024a) improves robustness by incorporating mean and standard deviation metrics across the full vocabulary. State-of-the-art (SOTA) methods like RECALL (Xie et al. 2024) and CON-RECALL (Wang et al. 2024) further improve detection by introducing dedicated prefixes. Both methods measure the relative shift in a model's log-likelihood for a sample when conditioned on specific prefixes. The underlying hypothesis is that non-member prefixes trigger more significant log-likelihood changes on member samples than non-member samples (or vice versa). By utilizing this asymmetry, such scoring functions achieve effective divergence between member and non-member samples.

Despite their promising results, these prefix-based meth-

*Corresponding author

ods still face three critical limitations. First, their performance exhibits substantial instability across different prefixes. As presented in Figure 1, RECALL's AUC performance varies by up to 30% under different non-member prefixes, while CON-RECALL's performance fluctuates by as much as 70% with different pairs of member and non-member prefixes. Unfortunately, both methods rely on random selection, which often leads to suboptimal results. Second, these methods assume access to a batch of high-quality labeled member and non-member prefixes. However, obtaining such labeled data is often impractical when manual costs are constraints. Third, while these prefix-based methods are grounded in intuitions and empirically validated, they lack theoretical analysis to explain the underlying principles.

To address these limitations, we first conduct a theoretical analysis to assess the impact of prefixes on existing prefix-based methods. Specifically, we conceptualize prefix injection as a type of in-context learning (ICL) and reveal it as forms of implicit fine-tuning (Dai et al. 2022). Our analysis demonstrates that the impact of prefixes can be measured via the cosine similarity between the gradient of the prefix and the gradient of the input sample. Optimal prefixes are those that maximize the divergence between the similarity score distributions of member and non-member samples.

However, as acquiring gradient information from LLMs is often impractical, we approximate the cosine similarity (between the gradient of a prefix and the gradient of an input sample) using S_{recall}, a metric representing the relative change in an input sample's log-likelihood when conditioned on a prefix. Thus, effective prefixes amplify distinctions in S_{recall} distributions between member and non-member data. To identify such prefixes, we propose a Discriminability Symmetry Criteria-guided Prefix Selection (DSC-Prefix) method, which derives two key criteria *Discriminability* and *Symmetry* from the S_{recall} distribution. These criteria serve to quantify the effectiveness of prefixes in detecting pre-training data, enabling precise selection of high-performing candidate prefixes. We evaluate our method on the WikiMIA benchmark (Shi et al. 2023). Results reveal that our method boosts both the performance and robustness of RECALL and CON-RECALL. Additional experiments corroborate the effectiveness of our method.

Our contributions are summarized as follows:

- We conduct a theoretical analysis to assess the impact of prefixes on existing prefix-based methods (e.g., RECALL and CON-RECALL), laying the theoretical foundation for prefix optimization.
- We introduce a novel prefix selection framework, namely DSC-Prefix, which eliminates the reliance on random selection or manual curation of individual prefixes.
- Extensive experiments on the WikiMIA dataset demonstrate that our method consistently boosts RECALL and CON-RECALL, achieving gains of up to 21.1% in AUC scores while significantly enhancing robustness.

Related Work

Membership Inference Attacks (MIAs) determine whether specific data samples were used to train a model. Proposed by Shokri et al. (2017), MIAs assess model memorization, privacy risks, and potential test data contamination (Carlini et al. 2022; Mireshghallah et al. 2022; Steinke, Nasr, and Jagielski 2023; Oren et al. 2023). Applied to pretrained LLMs, this is called pre-training data detection. While widely studied (Watson et al. 2021; Fu et al. 2023; Maini et al. 2024; Kandpal et al. 2024), MIAs on pretrained LLMs face challenges: significantly larger datasets and fewer training epochs (Carlini et al. 2022; Shi et al. 2023), blurring the distinction between member and non-member samples.

To address the pre-training data detection task, Min-K% (Shi et al. 2023) calculates membership scores by averaging the log-likelihood of the lowest-probability tokens (k% threshold). Min-K%++ (Zhang et al. 2024a) enhances robustness by incorporating mean and standard deviation across the full vocabulary. TAG&TAB (Antebi et al. 2025) combines keyword extraction with average log-likelihood scoring. DC-PDD (Zhang et al. 2024b) measures divergence between the model's token probabilities and reference frequencies for membership inference. In contrast, state-of-the-art (SOTA) methods use prefix-based strategies: RECALL (Xie et al. 2024) infers membership by analyzing log-likelihood shifts when prefixed with non-member context, while CON-RECALL (Wang et al. 2024) improves inference accuracy by comparing shifts under both member and non-member contexts.

However, the effectiveness of existing prefix-based methods heavily relies on specific prefixes. Random sampling, used in these approaches, often leads to performance instability. Our work differs from previous approaches by introducing an unsupervised prefix selection framework to improve both detection accuracy and stability.

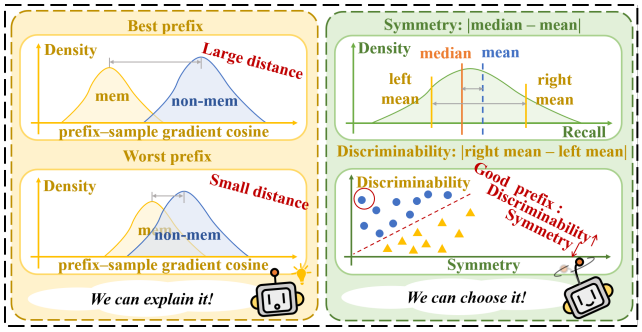

Figure 2: Overview of our prefix selection framework.

Methodology

We begin by providing a formal description of the pre-training data detection task. Next, we introduce preliminaries by reviewing the RECALL and CON-RECALL methods. We then perform a theoretical analysis, backed by quantitative evidence, to motivate our approach. Finally, we introduce a novel prefix selection method. The overall framework of our method is illustrated in Figure 2.

Problem Definition

Consider a model $\mathcal{M}$ pre-trained on a dataset $\mathcal{D}$. The goal of pre-training data detection is to determine whether a data point x belongs to the pre-training dataset $\mathcal{D}$, i.e., determining if $x \in \mathcal{D}$ (member sample) or $x \notin \mathcal{D}$ (non-member sample). To achieve this we develop a score function $s(x, \mathcal{M}) \to \mathbb{R}$ that assigns a scalar value to x based on model $\mathcal{M}$'s output. This score is then compared against a predefined threshold τ to determine selection:

$$\begin{cases} x \in \mathcal{D} & \text{if } s(x, \mathcal{M}) \geq \tau \\ x \notin \mathcal{D} & \text{if } s(x, \mathcal{M}) < \tau \end{cases} \tag{1}$$

Under the gray-box setting, detection methods can only access the model's output (e.g., logits and probabilities). Internal components, such as architectural parameters, intermediate layer activations, or gradient information, are not available.

Preliminaries

RECALL RECALL (Xie et al. 2024) quantifies the relative change in a model's log-likelihood for data point x when conditioned on a non-member prefix ($P_{\text{non-mem}}$). Formally, the scoring function $S_{\text{recall}}(x, \mathcal{M})$ is defined as:

$$S_{\text{recall}}(x, \mathcal{M}) = \frac{LL(x|P_{\text{non-mem}})}{LL(x)} \tag{2}$$

where $LL(x|P_{\text{non-mem}})$ is the conditional log-likelihood of x given the non-member prefix, and $LL(x)$ is the unconditional log-likelihood.

The core hypothesis is that member samples generally exhibit higher S_{recall} scores than non-member samples, allowing for discrimination between the two classes.

CON-RECALL CON-RECALL (Wang et al. 2024) extends RECALL by introducing contrastive conditioning through both non-member ($P_{\text{non-mem}}$) and member (P_{mem}) prefixes. Its scoring function $S_{\text{con-recall}}(x, \mathcal{M})$ is defined as:

$$S_{\text{con-recall}}(x, \mathcal{M}) = \frac{LL(x|P_{\text{non-mem}}) - \gamma \cdot LL(x|P_{\text{mem}})}{LL(x)} \tag{3}$$

where hyperparameter γ controls the contrast strength.

This function builds on the observation that non-member prefixes have a more pronounced effect on member samples' log-likelihoods compared to non-member samples. Conversely, member prefixes exert a greater influence on non-member samples' log-likelihoods. This asymmetry is modeled to amplify the discrimination.

Analysis on Prefix Influence

Our analysis builds on the theoretical framework of Dai et al. (2022), which interprets in-context learning (ICL) as a specialized form of implicit fine-tuning. Consider W_0 as the initial model parameters. After introducing context, the parameters update to $W_1 = W_0 + \Delta W_{\text{ICL}}$, where ΔW_{ICL} denotes the parameter adjustment induced by the context.

We conceptualize prefix P_α injection as ICL. Let $W^{(\alpha)} = W_0 + \Delta W_{\text{ICL}}^{(\alpha)}$ represent the updated parameters after incorporating prefix P_α, where $\Delta W_{\text{ICL}}^{(\alpha)}$ denotes the implicit parameter changes. In zero-shot inference, the model's unconditional log-likelihood for input x is $LL(x) = LL(x; W_0)$. When conditioned on prefix P_α, this becomes $LL(x|P_\alpha) = LL(x; W^{(\alpha)})$. Expanding $LL(x; W^{(\alpha)})$ via Taylor series around W_0 yields:

$$\begin{aligned} LL(x; W^{(\alpha)}) &= LL(x; W_0) + \nabla_W LL(x; W_0) \cdot \Delta W_{\text{ICL}}^{(\alpha)} \\ &+ \frac{1}{2} \left[\Delta W_{\text{ICL}}^{(\alpha)}\right]^\top \cdot \nabla_W^2 LL(x; W_0) \cdot \Delta W_{\text{ICL}}^{(\alpha)} + o\Big(||\Delta W_{\text{ICL}}^{(\alpha)}||^2\Big) \end{aligned} \tag{4}$$

Given the vast scale of LLM pre-training data and the stochastic nature of batch sampling for updates (Carlini et al. 2022; Shi et al. 2023), the actual gradient updates for a sample remains quite limited. Thus, in Eq. (4), the first-order term dominates in the Taylor series.[1] We approximate:

$$LL(x; W^{(\alpha)}) \approx LL(x; W_0) + \nabla_W LL(x; W_0) \cdot \Delta W_{\text{ICL}}^{(\alpha)} \tag{5}$$

Combining with Eq. (5), $S_{\text{recall}}(x, \mathcal{M})$ is simplified as:

$$\begin{aligned} S_{\text{recall}}(x, \mathcal{M}) &= \frac{LL(x|P_{\text{non-mem}})}{LL(x)} = \frac{LL(x; W^{(\text{non-mem})})}{LL(x; W_0)} \\ &\approx 1 + \frac{\nabla_W LL(x; W_0) \cdot \Delta W_{\text{ICL}}^{(\text{non-mem})}}{LL(x; W_0)} \end{aligned} \tag{6}$$

Similarly, $S_{\text{con-recall}}(x, \mathcal{M})$ is simplified as:

$$\begin{aligned} S_{\text{con-recall}}(x, \mathcal{M}) &= \frac{LL(x|P_{\text{non-mem}}) - \gamma \cdot LL(x|P_{\text{mem}})}{LL(x)} \\ &= \frac{LL(x; W^{(\text{non-mem})}) - \gamma \cdot LL(x; W^{(\text{mem})})}{LL(x; W_0)} \\ &\approx (1 - \gamma) + \frac{\nabla_W LL(x; W_0) \cdot \left[\Delta W_{\text{ICL}}^{(\text{non-mem})} - \gamma \cdot \Delta W_{\text{ICL}}^{(\text{mem})}\right]}{LL(x; W_0)} \end{aligned} \tag{7}$$

The computation of S_{recall}/$S_{\text{con-recall}}$ depends on three components: the base log-likelihood $LL(x; W_0)$, the gradient $\nabla_W LL(x; W_0)$ and the parameter changes ΔW_{ICL}.[2] The first two components derive from model's initial parameters and input sample itself. Therefore, prefixes influence S_{recall}/$S_{\text{con-recall}}$ through ΔW_{ICL}. Optimal prefixes are expected to maximize ΔW_{ICL}'s directional similarity with gradients of member samples while diverging from gradients of non-member samples (or vice versa), thus amplifying the score discrepancies between the two classes.

Quantitative Analysis For each sample x, we examine the cosine similarity $S_P(x)$ between the prefix-induced parameter changes ΔW_{ICL} and the sample's initial gradient $\nabla_W LL(x; W_0)$, defined as follows:

$$\begin{aligned} S_P(x) &= \cos(\Delta W_{\text{ICL}}, \nabla_W LL(x; W_0)) \\ &= \frac{\langle \Delta W_{\text{ICL}}, \nabla_W LL(x; W_0) \rangle}{\|\Delta W_{\text{ICL}}\| \|\nabla_W LL(x; W_0)\|} \end{aligned} \tag{8}$$

where $\cos(\cdot)$ denotes the cosine similarity function.

Since ΔW_{ICL} is not directly measurable, we approximate it using the gradients of the prefixes (i.e. $\nabla_W LL(P; W_0)$)

[1] Please refer to the Appendix for detailed evidence supporting this approximation.

[2] For simplicity, we use ΔW_{ICL} to denote both $\Delta W_{\text{ICL}}^{(\text{mem})}$ and $[\Delta W_{\text{ICL}}^{(\text{non-mem})} - \gamma \cdot \Delta W_{\text{ICL}}^{(\text{mem})}]$.

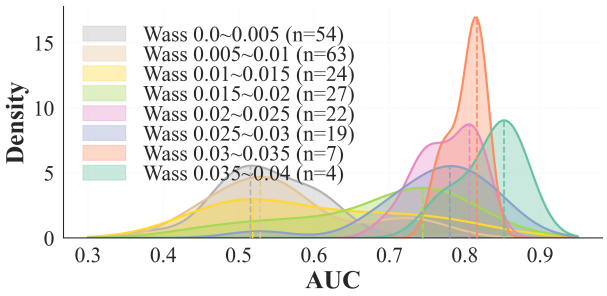

Figure 3: AUC distributions for different Wasserstein distance intervals.

in practice. The $S_P(x)$ sets for member and non-member samples are defined as:

$$\mathbb{S}_{\text{mem}} = \{S_P(x) | x \in \mathcal{D}\}, \quad \mathbb{S}_{\text{non-mem}} = \{S_P(x) | x \notin \mathcal{D}\} \tag{9}$$

An effective prefix or prefix pair should maximize separation between $\mathbb{S}_{\text{mem}}$ and $\mathbb{S}_{\text{non-mem}}$.

To this end, we use the Wasserstein distance (Dobrushin 1970) to measure the difference between $\mathbb{S}_{\text{mem}}$ and $\mathbb{S}_{\text{non-mem}}$ across all prefixes. A larger distance indicates a more pronounced difference in distributions. Figure 3 shows that a larger Wasserstein distance corresponds to a more concentrated AUC distribution in higher value intervals.

Prefix Selection Method

Building on these analyses, the core of selecting prefixes lies in maximizing the divergence between $\mathbb{S}_{\text{mem}}$ and $\mathbb{S}_{\text{non-mem}}$. Since this study treats LLMs as gray-box models without gradient access, we approximate the similarity score $S_P(x)$ using S_{recall}. We do this because $LL(x; W_0)$ in Eq. (6) provides little discriminative power (see Loss results in Table 1), while RECALL's key utility stems from $\nabla_W LL(x; W_0) \cdot \Delta W_{\text{ICL}}^{(\text{non-mem})}$, making S_{recall} a natural proxy.

For a prefix P, we first calculate S_{recall} across all test samples. Prefix selection then relies on two criteria derived from the S_{recall} distribution:

- *Discriminability*: The median of the S_{recall} distribution partitions the scores into inferred member and non-member groups. *Discriminability* is then defined as the distance between the respective means of these two groups.
- *Symmetry*: This measures the difference between the median and mean of the S_{recall} distribution. A small *Symmetry* prevents large *Discriminability* from outliers.

Formally, *Discriminability* (Dis_P) and *Symmetry* (Sym_P) for a prefix P are defined as:

$$\text{Dis}_P = |\mu_P^+ - \mu_P^-|, \quad \text{Sym}_P = |m_P - \mu_P| \tag{10}$$

where m_P and μ_P denote the median and mean of the S_{recall} distribution, respectively. μ_P^+ and μ_P^- denote the means of two sub-distributions split by the median.

Our method, namely Discriminability Symmetry Criteria-guided Prefix Selection (DSC-Prefix), prioritizes maximizing *discriminability* while minimizing *symmetry* to ensure robust separation between member and non-member distributions.

Leveraging DSC-Prefix for RECALL We employ predefined thresholds θ_{Sym} and θ_{Dis} to select prefixes. For any prefix P, it is regarded effective and added to the candidate set if it satisfies:

$$\text{Sym}_P < \theta_{\text{Sym}}, \quad \text{Dis}_P > \theta_{\text{Dis}} \tag{11}$$

RECALL then samples prefixes from this candidate set for inference.

Leveraging DSC-Prefix for CON-RECALL For CON-RECALL, we select prefix pairs (P_0, P_1), where P_0 follows the criteria in Eq. (11). P_1 is chosen to amplify the discrepancy in score distribution the two classes. This requires complementary thresholds θ_{Sym}^* and θ_{Dis}^*, with P_1 satisfying:

$$\text{Sym}_{P_1} < \theta_{\text{Sym}}^*, \quad \text{Dis}_{P_1} < \theta_{\text{Dis}}^* \tag{12}$$

P_1 meeting the conditions is added to a secondary candidate set. CON-RECALL then samples P_0 and P_1 from their respective candidate sets to form pair (P_0, P_1).

Experiment

Setup

Implementation details For prefix-based approaches, prefixes are sampled from the test dataset. To address the inherent randomness, we report the mean and standard deviation of AUC scores across 10 independent runs. To verify the effectiveness of our method, we adopt three selection strategies for comparison:

- **Unlabeled method**: Prefixes are randomly sampled without access to membership labels.
- **Labeled method**: Membership labels are accessible. Prefixes are randomly sampled from subsets (member or non-member). This method mirrors original implementations of RECALL (requires manual labeling of non-member prefixes) and CON-RECALL (relies on labeled member and non-member prefixes).
- **Our method**: Generate candidate sets for prefixes based our proposed criteria, then perform random sampling from the candidate sets.

For our method's candidate set generation, we sort prefixes by ascending *Discriminability* and *Symmetry* scores. Threshold parameters θ_{Sym}, θ_{Dis}, θ_{Sym}^* and θ_{Dis}^* are set as percentile proportions at 30%, 70%, 10% and 10%, respectively. Due to space limitation, other setups including "Dataset", "Evaluation Metrics", "Baseline Descriptions" and "LLM Configurations" are presented in Appendix.

Main Results

Table 1 compares our method with prior baselines. We make three key observations.

First, RECALL‡ and CON-RECALL‡ show performance (AUC and TPR@5%FPR) similar to early baselines (Loss, Ref, Zlib, Min-K%). This indicates that random sampling struggles to pick effective prefixes without access to membership labels.

Second, RECALL§ and CON-RECALL§ consistently outperform RECALL‡ and CON-RECALL‡ across all sub-datasets and LLMs. This suggests that both methods rely on

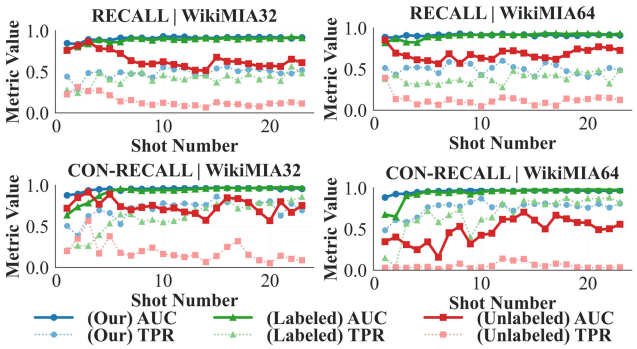

Figure 4: Performance comparison of three prefix selection methods—Unlabeled, Labeled and Ours—evaluated under varying prefix shot using the Pythia-6.9B model.

specific types of prefixes to achieve good results, highlighting the importance of prefix selection.

Third, RECALL* and CON-RECALL* shows consistent improvements over RECALL§ and CON-RECALL§. For example, on the length-32 sub-dataset, RECALL* improves AUC by 5.12 and TPR@5%FPR by 6.30 in average while CON-RECALL* improves AUC by 7.62 and TPR@5%FPR by 11.68 in average. Similar improvements can be observed across other sub-datasets. Overall, our method also has lower standard deviation in 70% of cases, showing that our selection strategy is more stable.

Performance on Few-Shot Prefixes

Employing multiple concatenated prefixes as input can further boost performance compared to a single prefix. In this section, we evaluate our method under different shots of prefixes. Figure 4 illustrates the AUC and TPR@5%FPR results for RECALL and CON-RECALL across three prefix selection strategies; Unlabeled, Labeled, and Ours, on WikiMIA-32 and WikiMIA-64 datasets.[3]

First, consider the AUC curves. Notably, even with very few shots, our method outperforms the Labeled strategy across all datasets. As the number of prefixes increases, the performance of our method and the Labeled strategy converges to comparable levels. Both approaches consistently surpass the Unlabeled strategy, highlighting the critical role of prefix selection in few-shot settings.

Examining the TPR@5%FPR curves reveals similar trends: our method exceeds both the Labeled and Unlabeled strategies across most of shot number. This further demonstrates its robustness with few-shot prefixes, reinforcing the advantages of our approach.

Further Analysis

This section provides an in-depth analysis of our prefix selection method. Due to space limitation, we focus on the RECALL method evaluated with the Pythia-6.9B model. Detailed analysis of CON-RECALL and results for alternative LLMs are provided in the appendix.

[3]Results for WikiMIA-128 and WikiMIA-256 are included in the Appendix due to space constraints.

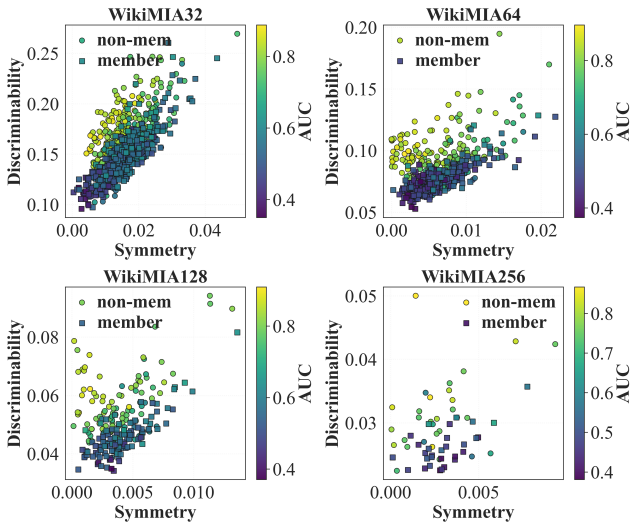

Figure 5: Scatter plots showing the distributions of *Symmetry* (Sym) and *Discriminability* (Dis) across prefixes, with color intensity indicating AUC score.

Effectiveness of *Discriminability* and *Symmetry*

This section examines how *Discriminability* (Dis) and *Symmetry* (Sym) metrics influence prefix selection. To visualize the relationship between the two metrics and detection performance, we plot all prefixes' Sym values on the x-axis and Dis values on the y-axis, with color intensity representing their corresponding AUC scores. Figure 5 illustrates these relationships across four sub-datasets.

Prefixes with low Sym values and high Dis values generally achieve superior AUC performance. Conversely, prefixes with both low Sym and low Dis exhibit poor discriminative power, leading to the weak AUC performance. Similar phenomenons can be observed across all four sub-datasets. These findings demonstrate two insights: First, our prefix selection method effectively identifies high-performance prefixes. Second, non-member prefixes generally induce higher AUC scores than member prefixes.

To investigate this performance gap between the member and non-member prefixes, we separately compute average cosine similarity within each class of samples (non-member vs. member), as shown in Table 2. Non-member samples exhibit significantly higher gradient direction uniformity compared to the more dispersed gradients of member samples. This suggests that perturbing inputs with member prefixes introduces greater variability in S_{recall} distributions within the member class, complicating detection efforts.

Performance Comparison under Sub-datasets

For each prefix, our method involves computing the S_{recall} distribution across all test samples. This entails high computational complexity, which escalates significantly as data size grows.

To enhance efficiency, we explore the feasibility of performing prefix selection on subsets of the test dataset. Specifically, subsets are constructed by randomly sampling

Len	Method	Mamba-1.4B AUC	Mamba-1.4B TPR @5% FPR	Pythia-6.9B AUC	Pythia-6.9B TPR @5% FPR	LLaMA-13B AUC	LLaMA-13B TPR @5% FPR	NeoX-20B AUC	NeoX-20B TPR @5% FPR	LLaMA-30B AUC	LLaMA-30B TPR @5% FPR	Average AUC	Average TPR @5% FPR
32	Loss	61.4	14.3	64.1	15.3	68.1	16.3	68.6	18.1	70.4	14.8	66.52	15.76
	Ref	49.3	4.7	59.6	8.3	56.1	9.3	72.0	13.7	58.9	10.3	59.18	9.26
	Zlib	62.5	13.5	64.3	13.7	68.4	15.5	68.6	19.7	70.6	16.3	66.88	15.74
	Min-K%	62.0	14.0	64.8	17.4	68.4	16.6	69.7	18.9	70.7	15.0	67.12	16.38
	Min-K%++	66.0	8.3	67.8	17.1	82.4	34.7	70.8	13.7	81.1	20.7	73.62	18.90
	Recall‡	$64.0_{\pm 13.5}$	$18.1_{\pm 10.6}$	$66.0_{\pm 11.1}$	$20.9_{\pm 9.1}$	$66.9_{\pm 7.4}$	$18.5_{\pm 5.4}$	$62.5_{\pm 13.4}$	$16.7_{\pm 10.4}$	$67.8_{\pm 7.0}$	$18.5_{\pm 8.7}$	65.44	18.54
	Recall§	$75.0_{\pm 6.8}$	$26.4_{\pm 8.6}$	$77.2_{\pm 6.3}$	$30.4_{\pm 8.7}$	$77.6_{\pm 6.2}$	$28.1_{\pm 5.7}$	$77.1_{\pm 5.5}$	$28.5_{\pm 8.8}$	$81.1_{\pm 6.2}$	$32.7_{\pm 8.3}$	77.60	29.22
	Recall*	$81.8_{\pm 4.4}$	$35.5_{\pm 9.8}$	$82.7_{\pm 2.2}$	$37.8_{\pm 7.5}$	$81.8_{\pm 2.8}$	$33.9_{\pm 6.5}$	$83.4_{\pm 2.3}$	$35.6_{\pm 5.3}$	$83.9_{\pm 3.3}$	$34.8_{\pm 8.4}$	82.72	35.52
	Con-recall‡	$60.9_{\pm 17.5}$	$15.7_{\pm 12.8}$	$62.9_{\pm 17.3}$	$18.4_{\pm 11.6}$	$68.7_{\pm 13.5}$	$19.1_{\pm 10.5}$	$64.4_{\pm 15.3}$	$20.9_{\pm 9.6}$	$70.3_{\pm 13.6}$	$21.4_{\pm 11.4}$	65.44	19.10
	Con-recall§	$76.9_{\pm 5.0}$	$26.1_{\pm 8.4}$	$75.7_{\pm 10.2}$	$30.0_{\pm 14.5}$	$84.0_{\pm 6.5}$	**$39.6_{\pm 12.7}$**	$75.8_{\pm 8.2}$	$30.9_{\pm 11.4}$	$84.1_{\pm 6.8}$	$37.8_{\pm 12.9}$	79.30	32.88
	Con-recall*	**$88.4_{\pm 3.3}$**	**$50.3_{\pm 11.1}$**	**$87.1_{\pm 4.0}$**	**$46.2_{\pm 11.4}$**	**$85.1_{\pm 3.7}$**	$39.3_{\pm 9.5}$	**$85.9_{\pm 3.7}$**	**$42.4_{\pm 10.2}$**	**$88.1_{\pm 2.5}$**	**$44.6_{\pm 7.8}$**	**86.92**	**44.56**
64	Loss	55.6	8.2	58.3	10.5	62.6	11.7	63.3	14.4	67.8	13.2	61.52	11.60
	Ref	48.4	6.2	64.3	8.2	59.4	16.3	74.9	22.2	61.7	13.2	61.74	13.22
	Zlib	58.1	14.4	60.5	15.2	64.5	12.5	65.4	18.7	66.6	16.3	63.02	15.42
	Min-K%	56.1	7.4	59.1	10.1	62.9	11.3	64.7	16.0	65.0	13.6	61.56	11.68
	Min-K%++	62.9	6.6	64.1	12.8	77.6	24.5	67.1	12.8	74.6	19.1	69.26	15.16
	Recall‡	$58.6_{\pm 13.5}$	$11.9_{\pm 8.8}$	$69.1_{\pm 15.2}$	$23.5_{\pm 11.4}$	$70.3_{\pm 11.0}$	$22.5_{\pm 11.1}$	$62.4_{\pm 10.8}$	$16.5_{\pm 9.4}$	$72.5_{\pm 12.1}$	$27.5_{\pm 15.5}$	66.58	20.38
	Recall§	$71.1_{\pm 8.1}$	$19.7_{\pm 8.9}$	$75.5_{\pm 10.1}$	$28.4_{\pm 11.3}$	$80.5_{\pm 6.3}$	$34.1_{\pm 11.1}$	$76.0_{\pm 6.6}$	$28.6_{\pm 8.8}$	$84.2_{\pm 4.2}$	$37.4_{\pm 7.0}$	77.46	29.64
	Recall*	$84.4_{\pm 2.1}$	$33.4_{\pm 8.7}$	$84.9_{\pm 2.8}$	$38.9_{\pm 7.3}$	$79.3_{\pm 9.8}$	$33.8_{\pm 13.6}$	$85.3_{\pm 2.0}$	$40.7_{\pm 7.8}$	$83.5_{\pm 4.3}$	$36.8_{\pm 6.9}$	83.48	36.72
	Con-recall‡	$56.9_{\pm 17.6}$	$9.3_{\pm 9.7}$	$60.0_{\pm 17.1}$	$15.2_{\pm 13.4}$	$51.4_{\pm 16.1}$	$11.9_{\pm 12.3}$	$52.8_{\pm 15.3}$	$12.5_{\pm 9.5}$	$78.8_{\pm 13.2}$	$34.0_{\pm 14.8}$	59.98	16.58
	Con-recall§	$81.5_{\pm 4.7}$	$31.9_{\pm 13.4}$	$82.7_{\pm 4.2}$	$31.0_{\pm 11.4}$	**$86.0_{\pm 8.7}$**	$44.6_{\pm 13.0}$	$80.4_{\pm 8.9}$	$31.8_{\pm 11.9}$	**$88.2_{\pm 2.5}$**	**$47.6_{\pm 5.8}$**	83.76	37.38
	Con-recall*	**$87.3_{\pm 2.3}$**	**$38.7_{\pm 6.9}$**	**$86.9_{\pm 4.0}$**	**$45.6_{\pm 8.1}$**	$84.3_{\pm 10.8}$	**$45.2_{\pm 18.4}$**	**$90.0_{\pm 2.2}$**	**$51.7_{\pm 8.4}$**	$87.1_{\pm 4.4}$	$41.1_{\pm 9.2}$	**87.12**	**44.46**
128	Loss	64.2	14.6	65.8	15.5	68.4	17.3	70.3	19.1	70.2	18.2	67.78	16.94
	Ref	48.5	4.6	68.4	8.2	58.2	8.2	82.1	11.8	63.3	12.7	64.10	9.10
	Zlib	66.4	16.4	67.9	20.0	70.4	21.8	72.2	21.8	72.2	20.0	69.82	20.00
	Min-K%	65.2	11.8	66.7	17.3	68.5	18.2	71.9	19.1	70.5	19.1	68.56	17.10
	Min-K%++	63.9	5.5	65.0	11.8	74.5	30.0	71.4	24.6	73.6	18.2	69.68	18.02
	Recall‡	$62.8_{\pm 14.1}$	$19.5_{\pm 10.4}$	$67.5_{\pm 11.8}$	$20.0_{\pm 9.6}$	$67.9_{\pm 13.4}$	$25.4_{\pm 14.6}$	$68.6_{\pm 9.6}$	$19.9_{\pm 7.5}$	$74.1_{\pm 10.3}$	$26.4_{\pm 13.1}$	68.18	22.24
	Recall§	$80.3_{\pm 3.8}$	$33.8_{\pm 9.2}$	$79.1_{\pm 9.6}$	$34.0_{\pm 13.6}$	$80.0_{\pm 9.2}$	$35.9_{\pm 14.5}$	$80.0_{\pm 9.1}$	$31.1_{\pm 13.6}$	$84.4_{\pm 6.5}$	$37.7_{\pm 13.7}$	80.76	34.50
	Recall*	$82.9_{\pm 2.0}$	$36.5_{\pm 7.3}$	$86.5_{\pm 1.9}$	**$42.1_{\pm 9.7}$**	$85.3_{\pm 3.5}$	**$44.6_{\pm 8.4}$**	$83.4_{\pm 3.8}$	$35.7_{\pm 9.1}$	**$89.2_{\pm 1.8}$**	**$47.3_{\pm 5.2}$**	85.46	**41.24**
	Con-recall‡	$64.6_{\pm 8.4}$	$19.2_{\pm 9.4}$	$59.1_{\pm 20.7}$	$24.1_{\pm 18.8}$	$47.8_{\pm 11.1}$	$11.1_{\pm 5.0}$	$64.2_{\pm 10.7}$	$14.8_{\pm 6.5}$	$56.2_{\pm 22.3}$	$19.1_{\pm 21.0}$	58.38	17.66
	Con-recall§	$76.5_{\pm 10.4}$	$26.4_{\pm 12.3}$	$76.5_{\pm 8.0}$	$27.3_{\pm 8.7}$	**$87.6_{\pm 1.5}$**	$42.1_{\pm 7.4}$	$80.0_{\pm 8.3}$	$30.5_{\pm 13.3}$	$82.0_{\pm 6.4}$	$31.6_{\pm 5.2}$	80.52	31.58
	Con-recall*	**$86.4_{\pm 2.0}$**	**$39.4_{\pm 9.1}$**	**$86.5_{\pm 1.6}$**	$39.3_{\pm 4.8}$	$86.6_{\pm 3.2}$	$43.3_{\pm 7.9}$	**$86.0_{\pm 3.9}$**	**$36.5_{\pm 13.6}$**	$88.2_{\pm 3.8}$	$46.4_{\pm 11.5}$	**86.74**	40.98
256	Loss	67.3	6.7	70.8	6.7	72.7	20.0	73.9	16.7	74.3	13.3	71.80	12.68
	Ref	44.7	3.3	67.1	6.7	70.3	6.7	84.6	10.0	75.1	10.0	68.36	7.34
	Zlib	68.2	26.7	71.2	**33.3**	74.8	33.3	74.7	30.0	75.0	**36.7**	72.78	32.00
	Min-K%	69.6	10.0	72.6	6.7	73.3	23.3	75.9	20.0	74.0	13.3	73.08	14.66
	Min-K%++	64.0	10.0	61.9	16.7	78.6	26.7	70.5	20.0	75.1	13.3	70.02	17.34
	Recall‡	$58.2_{\pm 13.6}$	$15.8_{\pm 13.8}$	$58.4_{\pm 14.8}$	$14.7_{\pm 11.3}$	$73.4_{\pm 11.6}$	$30.0_{\pm 27.1}$	$66.4_{\pm 16.3}$	$16.7_{\pm 10.9}$	$76.5_{\pm 14.8}$	$30.0_{\pm 20.0}$	66.58	21.44
	Recall§	$75.4_{\pm 3.6}$	$23.3_{\pm 10.5}$	$76.7_{\pm 5.0}$	$24.7_{\pm 10.0}$	$90.0_{\pm 1.3}$	**$60.0_{\pm 9.7}$**	$81.1_{\pm 5.8}$	$22.2_{\pm 5.7}$	$88.0_{\pm 0.1}$	$25.0_{\pm 8.3}$	82.24	31.04
	Recall*	$79.6_{\pm 4.9}$	$39.2_{\pm 9.8}$	$75.6_{\pm 9.6}$	$27.3_{\pm 10.2}$	$90.0_{\pm 1.9}$	$54.2_{\pm 14.2}$	$85.0_{\pm 4.2}$	$26.7_{\pm 13.6}$	$91.3_{\pm 0.1}$	$31.7_{\pm 11.7}$	84.30	35.82
	Con-recall‡	$46.3_{\pm 6.0}$	$12.2_{\pm 1.6}$	$79.4_{\pm 1.6}$	$24.4_{\pm 9.6}$	$59.2_{\pm 9.4}$	$18.9_{\pm 12.9}$	$67.3_{\pm 9.4}$	$21.7_{\pm 5.0}$	$36.1_{\pm 10.6}$	$1.7_{\pm 1.7}$	57.66	15.78
	Con-recall§	$62.1_{\pm 6.8}$	$14.4_{\pm 9.6}$	$63.8_{\pm 10.0}$	$30.0_{\pm 12.5}$	$69.8_{\pm 11.9}$	$21.1_{\pm 11.3}$	$85.2_{\pm 5.5}$	$33.3_{\pm 23.3}$	$86.5_{\pm 1.3}$	$33.3_{\pm 3.3}$	73.48	26.42
	Con-recall*	**$86.3_{\pm 2.6}$**	**$56.7_{\pm 4.7}$**	**$80.0_{\pm 7.0}$**	$27.8_{\pm 18.1}$	**$90.9_{\pm 1.7}$**	$33.3_{\pm 21.3}$	**$91.0_{\pm 2.4}$**	**$58.3_{\pm 5.0}$**	**$92.3_{\pm 0.1}$**	$30.0_{\pm 6.7}$	**88.10**	**41.22**

Table 1: AUC and TPR@5%FPR performance comparisons under one-shot prefix. Best performances are marked in bold. ‡: Unlabeled method; §: Labeled method; *: Our method.

	Non-Member	Member
Average Cosine Similarity	0.0398±0.0399	0.0214±0.0348

Table 2: Consistency of gradient direction within each class of samples (non-member vs. member).

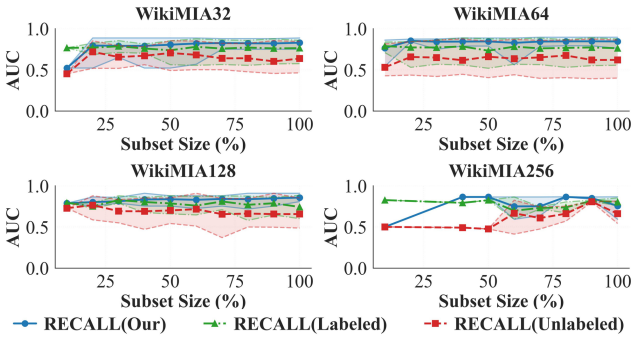

Figure 6: Performance under varying subsample ratios.

10%, 20%, 30%, ..., up to 100% of the test samples. The three prefix selection methods—Unlabeled, Labeled, and Ours—are applied to these subsets, and the selected prefixes are evaluated on the full test dataset.

As shown in Figure 6, subsets comprising just 20% of samples yield prefixes with good performance, substantially reducing computational costs. Notably, our method consistently surpasses both Unlabeled and Labeled approaches across all sampling ratios afterwards. For datasets with limited sample number, such as WikiMIA-256, higher subsample ratios are required to ensure effective prefix selection.

Performance on Imbalanced Class Distribution

Our method relies on *Discriminability* and *Symmetry* to select prefixes, which implicitly assumes a relatively balanced number of member and non-member samples. However, in many practical applications, there is often a pronounced imbalance between these groups. Thus, we further investigate the performance of our method under imbalanced membership distributions. Specifically, we construct test datasets with varying non-member/member ratios.

Figure 7 shows that as class ratios become increasingly imbalanced, the Unlabeled method exhibits the most instability, with overall lower AUC values and rapidly increasing fluctuations as imbalance intensifies. Our method outperforms the Labeled method when the dataset is relatively balanced. However, as the non-member to member ratio is extremely low (e.g., 1:9), our method underperforms the Labeled method. This is because, with too few non-member samples, our method struggle to identify high-quality non-member prefixes, whereas labeled methods leverage label information to accurately locate non-member prefixes, thus achieving superior performance.

Sensitivity Analysis on Quantile Proportions

We set thresholds θ_{Sym} and θ_{Dis} based on the two quantile proportions. To investigate our method's sensitivity to different quantile proportions, we fix the sum of the two quantile proportions at 100%, defining the x-axis as the proportion assigned to θ_{Sym}. Figure 8 presents the performances under varying quantile proportions.

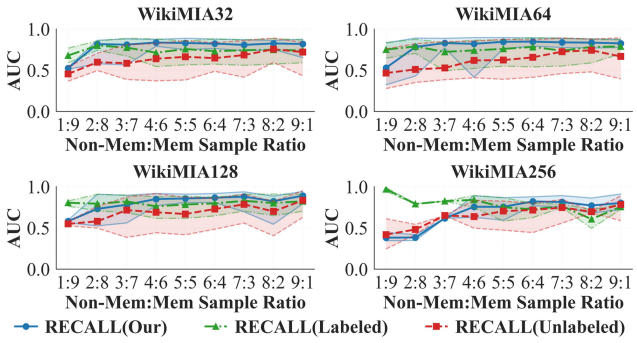

Figure 7: Performance under varying non-member/member ratios.

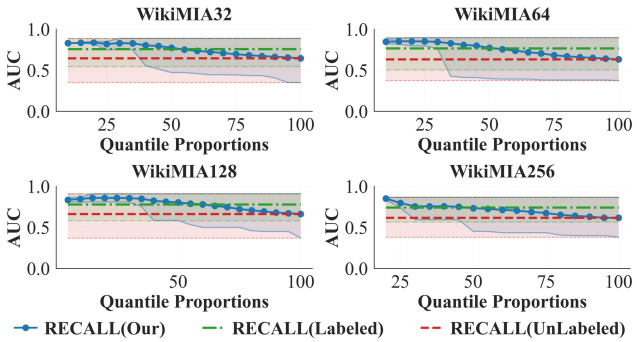

Figure 8: Performance comparisons under different quantile proportions across the four sub-datasets.

The results show that when the quantile proportion are set low, our method outperforms both the Labeled and Unlabeled methods. As the quantile proportion increases, the number of prefixes considered increases. The performance of our method and UnLabeled method converges. Moreover, at lower quantile proportion, the performance range of our method is narrower than that of the Labeled method, demonstrating greater robustness under high-confidence filtering.

Conclusion

This study investigates pre-training data detection for large language models (LLMs). We theoretically analyze how prefixes impact existing prefix-based detection methods and introduce two criteria—*Discriminability* and *Symmetry*—to select prefixes without relying on membership labels or manual effort. Experiments on the WikiMIA dataset demonstrate our method's effectiveness and robustness. Further analysis reveals strong generalization across diverse prefix scenarios (e.g., varying shot of prefixes, imbalanced data, and hyperparameter ranges). Future work will extend evaluation to broader real-world applications.

Acknowledgments

This research was partially supported by the Key Research and Development Project in Shaanxi Province No. 2023GXLH-024, the National Natural Science Foundation of China No. 62406242, 62476215, 62302380, 62037001, 62137002 and 62192781.

References

Antebi, S.; Habler, E.; Shabtai, A.; and Elovici, Y. 2025. Tag&tab: Pretraining data detection in large language models using keyword-based membership inference attack. *arXiv preprint arXiv:2501.08454*.

Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; and Zhang, C. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Dai, D.; Sun, Y.; Dong, L.; Hao, Y.; Ma, S.; Sui, Z.; and Wei, F. 2022. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*.

Dobrushin, R. L. 1970. Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, 15(3): 458–486.

Duarte, A. V.; Zhao, X.; Oliveira, A. L.; and Li, L. 2024. De-cop: Detecting copyrighted content in language models training data. *arXiv preprint arXiv:2402.09910*.

Fu, W.; Wang, H.; Gao, C.; Liu, G.; Li, Y.; and Jiang, T. 2023. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. *arXiv preprint arXiv:2311.06062*.

Kandpal, N.; Pillutla, K.; Oprea, A.; Kairouz, P.; Choquette-Choo, C. A.; and Xu, Z. 2024. User Inference Attacks on Large Language Models. arXiv:2310.09266.

Lin, X.; Huang, Z.; Zhang, Z.; Zhou, J.; and Chen, E. 2025. Explore What LLM Does Not Know in Complex Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24585–24594.

Maini, P.; Jia, H.; Papernot, N.; and Dziedzic, A. 2024. LLM Dataset Inference: Did you train on my dataset? arXiv:2406.06443.

Meeus, M.; Jain, S.; Rei, M.; and de Montjoye, Y.-A. 2024. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, 2369–2385.

Miah, M. S. U.; Kabir, M. M.; Sarwar, T. B.; Safran, M.; Alfarhood, S.; and Mridha, M. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, 14(1): 9603.

Mireshghallah, F.; Goyal, K.; Uniyal, A.; Berg-Kirkpatrick, T.; and Shokri, R. 2022. Quantifying privacy risks of masked language models using membership inference attacks. *arXiv preprint arXiv:2203.03929*.

Oren, Y.; Meister, N.; Chatterji, N. S.; Ladhak, F.; and Hashimoto, T. 2023. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*.

Shi, W.; Ajith, A.; Xia, M.; Huang, Y.; Liu, D.; Blevins, T.; Chen, D.; and Zettlemoyer, L. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Steinke, T.; Nasr, M.; and Jagielski, M. 2023. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 36: 49268–49280.

Tang, X.; Shin, R.; Inan, H. A.; Manoel, A.; Mireshghallah, F.; Lin, Z.; Gopi, S.; Kulkarni, J.; and Sim, R. 2023. Privacy-preserving in-context learning with differentially private few-shot generation. *arXiv preprint arXiv:2309.11765*.

Wang, C.; Wang, Y.; Hooi, B.; Cai, Y.; Peng, N.; and Chang, K.-W. 2024. Con-recall: Detecting pre-training data in llms via contrastive decoding. *arXiv preprint arXiv:2409.03363*.

Watson, L.; Guo, C.; Cormode, G.; and Sablayrolles, A. 2021. On the importance of difficulty calibration in membership inference attacks. *arXiv preprint arXiv:2111.08440*.

Xie, R.; Wang, J.; Huang, R.; Zhang, M.; Ge, R.; Pei, J.; Gong, N. Z.; and Dhingra, B. 2024. Recall: Membership inference via relative conditional log-likelihoods. *arXiv preprint arXiv:2406.15968*.

Xu, H.; Sharaf, A.; Chen, Y.; Tan, W.; Shen, L.; Van Durme, B.; Murray, K.; and Kim, Y. J. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Zhang, J.; Sun, J.; Yeats, E.; Ouyang, Y.; Kuo, M.; Zhang, J.; Yang, H. F.; and Li, H. 2024a. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*.

Zhang, W.; Zhang, R.; Guo, J.; de Rijke, M.; Fan, Y.; and Cheng, X. 2024b. Pretraining data detection for large language models: A divergence-based calibration method. *arXiv preprint arXiv:2409.14781*.