Detecting Training Data of Large Language Models via Expectation Maximization

Anonymous ACL submission

Abstract

Membership inference attacks (MIAs) aim to determine whether a specific example was used to train a given language model. While prior work has explored prompt-based attacks such as ReCALL, these methods rely heavily on the assumption that using known non-members as prompts reliably suppresses the model's responses to non-member queries. We propose EM-MIA, a new membership inference approach that iteratively refines prefix effectiveness and membership scores using an expectation-maximization strategy without requiring labeled non-member examples. To support controlled evaluation, we introduce OLMoMIA, a benchmark that enables analysis of MIA robustness under systematically varied distributional overlap and difficulty. Experiments on WikiMIA and OLMoMIA show that EM-MIA outperforms existing baselines, particularly in settings with clear distributional separability. We highlight scenarios where EM-MIA succeeds in practical settings with partial distributional overlap, while failure cases expose fundamental limitations of current MIA methods under near-identical conditions. We will release our code and evaluation pipeline upon publication to encourage reproducible and robust MIA research.

1 Introduction

As large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023b) continue to advance in scale and capability, growing concerns have emerged regarding the provenance and transparency of their training data (Henderson et al., 2023; Liang et al., 2023). This issue is crucial in both research and real-world deployments, where uncertainty about what data a model has seen can lead to legal and ethical risks, such as privacy breaches (Staab et al., 2023; Kandpal et al., 2023), copyright infringement (Meeus et al., 2024c), and the leakage of sensitive or proprietary content (Chang et al., 2023).

Membership inference attacks (MIAs) offer a concrete framework for probing this issue by attempting to determine whether a specific example was included in a model's training corpus (Shokri et al., 2017; Carlini et al., 2022). By doing so, they enable auditing of model behavior and exposure, helping practitioners evaluate data contamination (Magar and Schwartz, 2022; Sainz et al., 2023, 2024) or compliance with data usage policies (Voigt and Von dem Bussche, 2017; Legislature, 2018). Despite their utility, MIAs on LLMs remain fundamentally challenging due to the massive size of pre-training corpora and the subtle boundary between memorization and generalization in natural language (Duan et al., 2024). Recent work has proposed prompt-based MIA techniques such as ReCALL (Xie et al., 2024), which assume that known non-members can serve as effective prompts for distinguishing members from non-members. However, we find that the effectiveness of such prompts is highly inconsistent and difficult to predict, motivating the need for a more adaptive approach that can account for variability in prompt effectiveness.

To address the limitations of approaches that rely on arbitrarily or randomly chosen prompts, we propose EM-MIA, a novel membership inference method that jointly refines prefix effectiveness and membership scores through an expectation-maximization procedure. Our approach is motivated by the observation that the usefulness of a prompt, defined as its ability to differentiate members from non-members, varies widely across examples and cannot be reliably determined in advance. Instead of relying on labeled non-members or assuming the quality of predefined prompts, EM-MIA uses the model's own responses to iteratively estimate which prefixes are informative and which examples are likely to be members. This interaction allows the model to bootstrap its predictions over both prompt selection and membership esti-

mation in a fully unsupervised manner. As a result, EM-MIA offers greater flexibility and robustness across diverse settings, particularly when prompt-based assumptions do not hold or ground-truth non-member data is unavailable.

To facilitate more controlled and reproducible evaluation of membership inference methods, we introduce OLMoMIA, a benchmark constructed from the pre-training corpus and checkpoints of the OLMo open-source LLM series (Groeneveld et al., 2024). Unlike existing benchmarks such as WikiMIA (Shi et al., 2023) and MIMIR (Duan et al., 2024), which provide limited control over the similarity between member and non-member examples, OLMoMIA allows researchers to systematically vary distributional overlap and assess how different methods perform across a range of difficulty levels. By partitioning the data based on semantic similarity and membership status with respect to the pre-training data, OLMoMIA supports fine-grained analysis of robustness, generalization, and failure modes in both easy and near-indistinguishable settings. Its design enables rigorous comparison of inference strategies under controlled conditions, and we will release both the benchmark and its generation pipeline to support scalable and reproducible MIA research.

Our experiments show that EM-MIA outperforms existing MIA methods on WikiMIA across models of varying sizes and achieves robust results on OLMoMIA under systematically controlled difficulty conditions. In particular, EM-MIA demonstrates strong performance without access to labeled non-member data and maintains robustness to prompt variability, highlighting its practical value in realistic gray-box scenarios. At the same time, our results expose the inherent difficulty of membership inference when member and non-member distributions are nearly identical, which poses a significant challenge for all existing methods, including ours. These findings underscore the importance of evaluating MIA methods across a range of separability conditions and offer new insight into the limits and opportunities of prompt-based membership inference.

2 Related Work

Membership Inference on LLMs. Membership inference on LLMs presents unique challenges. First, LLMs are trained on massive corpora, and individual examples are typically seen only once or a few times (Lee et al., 2021), leaving minimal memorization footprint. Second, defining membership is inherently ambiguous in natural language, in that texts often repeat or partially overlap even after rigorous decontamination (Kandpal et al., 2022; Tirumala et al., 2024), and paraphrased or semantically similar content can blur membership boundaries (Shilov et al., 2024; Mattern et al., 2023; Mozaffari and Marathe, 2024). Traditional MIA methods often rely on training shadow models using labeled data from a similar distribution (Shokri et al., 2017), but this is impractical in LLM settings due to limited access to comparable data and training specifications.

In contrast, MIA methods for LLMs typically use the model's loss (e.g., negative log-likelihood) as a membership score, under the assumption that models tend to memorize or overfit their training data (Yeom et al., 2018; Carlini et al., 2022). Building on this idea, several techniques calibrate membership scores based on input difficulty (Ye et al., 2022), using reference models (Carlini et al., 2022), compression-based heuristics (Carlini et al., 2021), or nearest neighbors in embedding space (Mattern et al., 2023). Other methods focus on low-likelihood tokens (Shi et al., 2023) or compute calibrated token-level ratios (Zhang et al., 2024).

ReCALL (Xie et al., 2024) proposes a different strategy by using known non-member examples as prompts to condition the model's response. It assumes that such prompts suppress memorization signals, enabling members to stand out by their elevated likelihood under the same prompt. However, this assumption is brittle, as prompt effectiveness varies significantly across examples, and a fixed prompt often fails to generalize across models or domains. We address this limitation by proposing a fully unsupervised method that jointly estimates prompt effectiveness and membership likelihood, without relying on labeled non-members or fixed prompting strategies.

Evaluation Benchmarks. Robust evaluation of MIA methods for LLMs remains challenging because existing benchmarks rarely provide both reliable membership labels and controllable distributional settings. Most benchmarks fall into one of two categories. Some, such as WikiMIA (Shi et al., 2023; Meeus et al., 2024a), determine membership based on document timestamps and model release dates. This approach risks conflating membership inference with distribution shift detection (Das

et al., 2024; Meeus et al., 2024b; Maini et al., 2024). Others, such as MIMIR (Duan et al., 2024), use random splits to ensure that member and non-member distributions are nearly identical. In such cases, no existing method performs significantly better than random guessing.

These limitations make it difficult to understand how well a method generalizes across different data conditions. Pre-training corpora are typically drawn from diverse sources, while inference-time inputs may come from entirely different domains. Effective evaluation therefore requires testing under a range of membership separability conditions. However, constructing such benchmarks is practically difficult, especially given the lack of true non-member data and the challenge of controlling test distributions. There is a clear need for evaluation setups that reflect varied, realistic scenarios while maintaining access to reliable ground-truth labels (Meeus et al., 2024b; Eichler et al., 2024).

3 Method

3.1 Problem Formulation

We consider membership inference in a gray-box setting, where the attacker has access to a language model $\mathcal{M}$ and can query $\mathcal{M}$ to obtain token-level probabilities or log-likelihoods. Given an input $x \in \mathcal{D}_{\text{test}}$, the goal is to predict a binary membership label indicating whether x was included in the pretraining corpus $\mathcal{D}_{\text{train}}$ of $\mathcal{M}$.

3.2 ReCaLL: Assumptions and Limitations

ReCaLL (Xie et al., 2024) is a prompt-based membership inference method that computes the ratio between the conditional and unconditional log-likelihoods of a target example x under $\mathcal{M}$. Given a prefix p, the ReCaLL score is defined as $\text{ReCaLL}_p(x; \mathcal{M}) = \text{LL}(x \mid p; \mathcal{M})/\text{LL}(x; \mathcal{M})$, where LL denotes the average log-likelihood over tokens, and $p = p_1 \oplus \cdots \oplus p_n$ is a concatenation of non-member examples p_i. The intuition is that conditioning on non-members tends to reduce the likelihood of members more than that of non-members, making the ratio indicative for membership prediction.

ReCaLL demonstrates strong empirical performance, achieving over 90% AUC-ROC on WikiMIA (Shi et al., 2023) and outperforming prior methods such as Min-K%++ (Zhang et al., 2024). However, this performance depends on strong assumptions and lacks theoretical justification. In its original implementation, ReCaLL constructs prefixes by randomly selecting non-members from the test set, assuming that (1) ground-truth non-members are available at inference time, and (2) all non-members are equally effective as prompts.

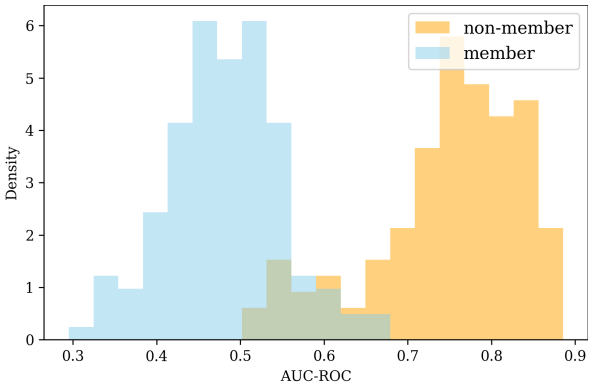


Figure 1: Distribution of prefix scores (measured by AUC-ROC in the oracle setting) for members and non-members on WikiMIA (Shi et al., 2023) (length 128) using Pythia-6.9B (Biderman et al., 2023).

In practice, such assumptions rarely hold so labeled non-members are often unavailable, especially when the training and test data distributions substantially overlap (Villalobos et al., 2022; Muennighoff et al., 2024). Even synthetic prefixes generated using GPT-4, as explored in Xie et al. (2024), rely on seed non-members drawn from the test distribution. This reliance on known non-members gives ReCaLL an unfair advantage over methods that operate without access to test labels.

Ablation studies in Xie et al. (2024) further show that ReCaLL's performance degrades when the prefix and test inputs differ in distribution, and that different random samples yield significant variance in accuracy. These findings suggest that non-members vary widely in their effectiveness as prompts, and that ReCaLL does not generalize reliably across domains or distribution shifts. These limitations motivate the need for a more flexible and fully unsupervised approach that does not depend on labeled non-members or assume prompt effectiveness in advance.

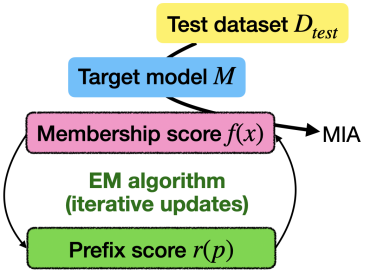
3.3 Motivation: Sensitivity to Prefix Choice

We empirically examine how ReCaLL's performance varies with the choice of prefix, particularly when labeled non-members are unavailable. To this end, we define a *prefix score* $r(p)$ as the effectiveness of a prefix p in distinguishing members from non-members when used in ReCaLL.

Algorithm 1 EM-MIA

Input: Target LLM $\mathcal{M}$, Test dataset $\mathcal{D}_{\text{test}}$
Output: Membership scores $f(x)$ for $x \in \mathcal{D}_{\text{test}}$

1: Initialize $f(x)$ with an existing off-the-shelf MIA method
2: **repeat**
3: Update prefix scores $r(p) = S(\text{ReCaLL}_p, f, \mathcal{D}_{\text{test}})$ for $p \in \mathcal{D}_{\text{test}}$
4: Update membership scores $f(x) = -r(x)$ for $x \in \mathcal{D}_{\text{test}}$
5: **until** Convergence (no significant difference in f)



In an oracle setting with access to ground-truth membership labels, we compute $r(p)$ as the AUC-ROC of $\text{ReCaLL}_p(x)$ over a test set $\mathcal{D}_{\text{test}}$, using each $x \in \mathcal{D}_{\text{test}}$ as a standalone prefix. This allows us to empirically measure the effectiveness of each test example when used as a prefix.

Figure 1 shows that non-member prefixes generally lead to strong ReCaLL performance, with AUC-ROC often exceeding 0.7. In contrast, member prefixes perform poorly, with scores clustering near 0.5 (i.e., random guessing). Additional comparisons using alternative metrics for prefix scoring are included in Appendix C. These results highlight two limitations of current ReCaLL-based methods: (1) Even among non-members, prefix effectiveness varies widely; (2) In realistic scenarios, ground-truth labels needed to evaluate or filter prefixes are unavailable.

These findings underscore the need for an approach that can identify effective prefixes and infer membership without access to labels. We address this challenge in the following section by proposing a fully unsupervised method that jointly estimates membership likelihood and prefix effectiveness through iterative refinement.

3.4 EM-MIA: Joint Estimation via EM

To address the practical setting where neither labeled non-members nor reliable prompt effectiveness can be assumed, we propose EM-MIA, a fully unsupervised method that jointly estimates prefix effectiveness and membership likelihood using an expectation-maximization (EM) procedure.

Let $f(x)$ denote the membership score for each test example $x \in \mathcal{D}_{\text{test}}$, and $r(p)$ denote the effectiveness score of a prefix p. The key insight is that membership scores and prefix scores can reinforce each other: better membership estimates allow more accurate estimation of prefix effectiveness, and more reliable prefixes lead to improved membership predictions. This mutual dependency motivates an iterative procedure in which each set of scores is refined based on the other.

Algorithm 1 outlines the overall procedure of EM-MIA. We initialize membership scores using any existing off-the-shelf MIA method such as Loss (Yeom et al., 2018) or Min-K%++ (Zhang et al., 2024) (Line 1). We then alternate between two updates: (1) estimating prefix scores $r(p)$ based on current membership scores $f(x)$ (Line 3), and (2) updating $f(x)$ using the refined $r(p)$ (Line 4). This process continues until convergence (Line 5). Because EM-MIA is a general framework, initialization, score update rules, stopping criteria, and datasets (see Appendix A) can be adapted to different applications.

Updating Prefix Scores. As shown in Section 3.3, AUC-ROC is an effective function S for evaluating a prefix p in the oracle setting given ground truth labels. Since ground-truth labels are not available, we generate pseudo-labels using a threshold τ over current membership scores $f(x)$ and use them to calculate prefix scores: $\text{AUC-ROC}(\{(\text{ReCaLL}_p(x), \mathbf{1}_{f(x)>\tau}) \mid x \in \mathcal{D}_{\text{test}}\})$. We typically set τ to the median of $f(x)$, assuming a balanced dataset. Alternatively, instead of relying on hard thresholds, we can measure rank alignment between $\text{ReCaLL}_p(x)$ and $f(x)$ using the average absolute rank difference or rank correlation coefficients such as Kendall's tau (Kendall, 1938) or Spearman's rho (Spearman, 1961).

Updating Membership Scores. Section 3.3 also shows that a negative prefix score $-r(x)$ is a simple yet effective membership score. Alternatively one could construct a prefix $p = p_1 \oplus \cdots \oplus p_n$ using top-k examples ranked by $r(x)$, and compute $f(x) = \text{ReCaLL}_p(x)$ using this prefix. The ordering of p_i within p is also a design choice. Placing stronger prefixes closer to x may amplify their influence due to LLMs' attention bias toward recent tokens.

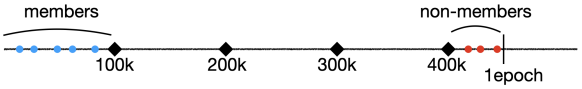


Figure 2: The basic setup of OLMoMIA benchmark. The horizontal line indicates a training step. For any intermediate checkpoint at a specific step, we can consider training data before and after that step as members and non-members, respectively.

4 OLMoMIA Benchmark

Motivation. To enable controlled and reproducible evaluation of MIA methods under varying difficulty levels, we introduce OLMoMIA, a new benchmark constructed from the training data and checkpoints of the OLMo-7B model (Groeneveld et al., 2024), which was pre-trained on the Dolma dataset (Soldaini et al., 2024). Unlike existing benchmarks such as WikiMIA (Shi et al., 2023), which rely on time-based heuristics, or MIMIR (Duan et al., 2024), which draws member and non-member examples from randomly partitioned subsets of the same data distribution, OLMoMIA allows systematic control over the distributional overlap between members and non-members. This allows evaluation under more realistic and ambiguous conditions, where membership inference is inherently more difficult.

Membership Label Assignment. Figure 2 illustrates the benchmark setup. OLMo provides intermediate model checkpoints and a detailed index mapping training steps to data examples, offering a rare opportunity to precisely define membership. We use four OLMo-7B checkpoints saved at 100k, 200k, 300k, and 400k training steps, where one full epoch consists of just over 450k steps. We define member examples as those seen before step 100k and non-members as those introduced between steps 400k and 500k. This setup reflects a practical incremental training scenario. Some ambiguity in membership may remain despite deduplication, as discussed in Section 2.

Dataset Sampling with Varying Difficulty We construct six dataset variants to simulate different levels of distributional overlap. The basic *Random* setting samples member and non-member examples uniformly from their respective intervals. This is analogous to MIMIR (Duan et al., 2024), which is known to be more challenging than WikiMIA due to minimal distributional differences between members and non-members (Gao et al., 2020).

To introduce controlled variation in difficulty, we first embed the candidate examples using NV-Embed-v2 (Lee et al., 2024), the top-performing model on the MTEB leaderboard (Muennighoff et al., 2022) as of August 2024. We then perform K-means clustering (Lloyd, 1982) separately on member and non-member embeddings with $K = 50$. To ensure diversity within clusters, we apply greedy deduplication by removing examples that are too similar (cosine distance below 0.6) to other points in the same cluster.

Based on these clusters, we define three difficulty-controlled variants: *Easy* selects the most dissimilar member and non-member clusters and samples examples furthest from the opposing group; *Hard* selects the most similar clusters and samples examples closest to the opposing group; *Medium* selects clusters with median inter-cluster distance and samples randomly from each.

We additionally define two hybrid settings: *Mix-1* combines members from *Random* and non-members from *Hard*, simulating tightly clustered test-time distributions; *Mix-2* does the reverse, combining members from *Hard* and non-members from *Random*. Together, these configurations span a broad range of separability conditions, providing a robust testbed for evaluating MIA methods. Formal definitions of each construction step are included in Appendix D.

Dataset Specifications. Each difficulty variant includes two subsets with maximum sequence lengths of 64 and 128 tokens. Each subset contains 500 members and 500 non-members, for a total of 1,000 examples per dataset.

Release Plan. We will release the OLMoMIA datasets along with the code used to generate each difficulty variant from the OLMo corpus and checkpoints. This will support scalable and reproducible MIA research under realistic gray-box conditions.

5 Experimental Setup

5.1 Datasets and Models

We evaluate EM-MIA and compare it with baseline methods on WikiMIA (§6.1) and OLMoMIA (§6.2) using AUC-ROC as a main evaluation metric. We also report TPR@1%FPR results in Appendix F. WikiMIA (Shi et al., 2023) provides length-based splits of 32, 64, and 128, and we follow prior work (Xie et al., 2024; Zhang et al., 2024) in using Mamba 1.4B (Gu and Dao, 2023), Pythia 6.9B (Bi-

derman et al., 2023), GPT-NeoX 20B (Black et al., 2022), LLaMA 13B/30B (Touvron et al., 2023a), and OPT 66B (Zhang et al., 2022) as target models. For OLMoMIA, we use all six controlled difficulty settings of *Easy*, *Medium*, *Hard*, *Random*, *Mix-1*, and *Mix-2*, and evaluate using OLMo-7B checkpoints after 100k, 200k, 300k, and 400k training steps. We exclude MIMIR (Duan et al., 2024) from our experiments since it lacks a baseline that performs meaningfully better than random guessing, which is required for initialization in EM-MIA.

5.2 Baselines

We compare EM-MIA against the following baselines: Loss (Yeom et al., 2018), Ref (Carlini et al., 2022), Zlib (Carlini et al., 2021), Min-K% (Shi et al., 2023), and Min-K%++ (Zhang et al., 2024). We use Pythia-70m for WikiMIA and StableLM-Base-Alpha-3B-v2 model (Tow, 2023) for OLMoMIA as the reference model of the Ref method, following Shi et al. (2023) and Duan et al. (2024). We use $K = 20$ for Min-K% and Min-K%++. Among the commonly used baselines, we omit Neighbor (Mattern et al., 2023) because it is not the best in most cases though it requires LLM inference multiple times for neighborhood texts, so it is much more expensive.

5.3 ReCaLL-based Baselines

We include several variants of ReCaLL that differ in how the prefix $p = p_1 \oplus \cdots \oplus p_n$ is constructed: *Rand*, *RandM*, *RandNM*, and *TopPref*. *Rand* randomly selects any data from $\mathcal{D}_{\text{test}}$. *RandM* randomly selects member data from $\mathcal{D}_{\text{test}}$. *RandNM* randomly selects non-member data from $\mathcal{D}_{\text{test}}$. *TopPref* selects data from $\mathcal{D}_{\text{test}}$ with the highest prefix scores calculated with ground truth labels the same as §3.3.

Among these, only *Rand*is fully unsupervised; the others either partially or fully rely on labels in the test dataset, making them unsuitable for realistic scenarios. For all methods using a random selection (*Rand*, *RandM*, and, *RandNM*), we execute five times with different random seeds and report the average. We fix $n = 12$ since it provides a reasonable performance while not too expensive. We report the results from the original ReCaLL paper but explain why this is not a fair comparison in Appendix B.

We also evaluate two unsupervised averaging variants. *Avg* and *AvgP* average ReCaLL scores over all data points in $\mathcal{D}_{\text{test}}$: $Avg(x) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{p \in \mathcal{D}_{\text{test}}} \text{ReCaLL}_p(x)$ and $AvgP(p) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{x \in \mathcal{D}_{\text{test}}} \text{ReCaLL}_p(x)$. The intuition is averaging will smooth out ReCaLL scores with a non-discriminative prefix while keeping ReCaLL scores with a discriminative prefix without exactly knowing discriminative prefixes.

5.4 EM-MIA

As described in Section 3.4, EM-MIA is a general framework where each component can be tuned for improvement, but we use the following options as defaults based on results from preliminary experiments. Overall, Min-K%++ performs best among baselines without ReCaLL-based approaches, so we use it as a default choice for initialization. Alternatively, we may use ReCaLL-based methods that do not rely on any labels like *Avg*, *AvgP*, or *Rand*. For the update rule for prefix scores, we use AUC-ROC as a default scoring function S. For the update rule for membership scores, we use negative prefix scores as new membership scores. For the stopping criterion, we repeat ten iterations and stop without thresholding by the score difference since we observed that membership scores and prefix scores converge quickly after a few iterations. We also observed that EM-MIA is not sensitive to the choice of the initialization method and the scoring function S and converges to similar results. Ablation study on different initializations and scoring functions can be found in Section 6.3. Discussion on computational costs can be found in Appendix E.

6 Results and Discussion

6.1 WikiMIA

Table 1 and Table 3 show results on WikiMIA, using AUC-ROC and TPR@1%FPR as evaluation metrics respectively. EM-MIA achieves state-of-the-art performance across all models and length splits, significantly outperforming all baselines, including ReCaLL, even without access to labeled non-member examples. In all cases, EM-MIA exceeds 96% AUC-ROC. For the largest model, OPT-66B, it reaches over 99% AUC-ROC for length 32 and 64, whereas ReCaLL falls below 86%.

All non-ReCaLL baselines remain below 76% AUC-ROC on average. The performance order among ReCaLL-based variants is consistent: *RandM* < *Avg*, *AvgP* < *Rand* < *RandNM* < *TopPref*. This pattern confirms that ReCaLL is highly sensitive to the choice of prefix. Particularly, the significant performance gap between *Rand* and *RandNM*

Method	Mamba-1.4B			Pythia-6.9B			LLaMA-13B			NeoX-20B			LLaMA-30B			OPT-66B			Average		
	32	64	128	32	64	128	32	64	128	32	64	128	32	64	128	32	64	128	32	64	128
Loss	61.0	58.2	63.3	63.8	60.8	65.1	67.5	63.6	67.7	69.1	66.6	70.8	69.4	66.1	70.3	65.7	62.3	65.5	66.1	62.9	67.1
Ref	60.3	59.7	59.7	63.2	62.3	63.0	64.0	62.5	64.1	68.2	67.8	68.9	65.1	64.8	66.8	63.9	62.9	62.7	64.1	63.3	64.2
Zlib	61.9	60.4	65.6	64.3	62.6	67.6	67.8	65.3	69.7	69.3	68.1	72.4	69.8	67.4	71.8	65.8	63.9	67.4	66.5	64.6	69.1
Min-K%	63.3	61.7	66.7	66.3	65.0	69.5	66.8	66.0	71.5	72.1	72.1	75.7	69.3	68.4	73.7	67.5	66.5	70.6	67.5	66.6	71.3
Min-K%++	66.4	67.2	67.7	70.2	71.8	69.8	84.4	84.3	83.8	75.1	76.4	75.5	84.3	84.2	82.8	69.7	69.8	71.1	75.0	75.6	75.1
Avg	70.2	68.3	65.6	69.3	68.2	66.7	77.2	77.3	74.6	71.4	72.0	68.7	79.8	81.0	79.6	64.6	65.6	60.0	72.1	72.1	69.2
AvgP	64.0	61.8	56.7	62.1	61.0	59.0	63.1	60.3	56.4	63.9	61.8	61.1	60.3	60.0	55.4	86.9	94.3	95.1	66.7	66.5	63.9
RandM	25.4	25.1	26.2	24.9	26.2	24.6	21.0	14.9	68.6	25.3	28.3	29.8	14.0	15.1	70.4	33.9	40.9	42.9	24.1	25.1	43.8
Rand	72.7	78.2	64.2	67.0	73.4	68.7	73.9	75.4	68.5	68.2	74.5	67.5	66.9	71.7	70.2	64.5	67.8	58.6	68.9	73.5	66.3
RandNM	90.7	90.6	88.4	87.3	90.0	88.9	92.1	93.4	68.8	85.9	89.9	86.3	90.6	92.1	71.8	78.7	77.6	67.8	87.5	88.9	78.7
TopPref	90.6	91.2	88.0	91.3	92.9	90.1	93.5	94.2	71.8	88.4	92.0	90.2	92.9	93.8	74.8	83.6	79.6	72.1	90.0	90.6	81.2
Xie et al. (2024)	90.2	91.4	91.2	91.6	93.0	92.6	92.2	95.2	92.5	90.5	93.2	91.7	90.7	94.9	91.2	85.1	79.9	81.0	90.1	91.3	90.0
EM-MIA	97.1	97.6	96.8	97.5	97.5	96.4	98.1	98.8	97.0	96.1	97.6	96.3	98.5	98.8	98.5	99.0	99.0	96.7	97.7	98.2	96.9

Table 1: AUC-ROC results on WikiMIA benchmark. The second block (grey) is ReCaLL-based baselines. *RandM*, *RandNM*, ReCaLL, and *TopPref* use labels in the test dataset, so comparing them with others is unfair. We report their scores for reference. We borrow the original ReCaLL results from Xie et al. (2024) which is also unfair to be compared with ours and other baselines.

Method	Easy		Medium		Hard		Random		Mix-1		Mix-2	
	64	128	64	128	64	128	64	128	64	128	64	128
Loss	32.5	63.3	58.9	49.0	43.3	51.5	51.2	52.3	65.7	49.0	30.8	54.7
Ref	56.8	26.8	61.4	47.2	49.1	50.7	49.7	49.9	59.9	49.7	38.9	50.9
Zlib	24.0	51.8	44.8	50.7	40.5	51.1	52.3	50.5	63.2	47.2	31.5	54.3
Min-K%	32.4	50.0	54.0	51.9	43.0	51.2	51.7	51.0	60.8	50.4	34.9	51.7
Min-K%++	45.2	59.4	56.4	45.7	46.4	51.4	51.0	51.9	57.9	50.0	39.8	53.2
Avg	61.9	53.9	52.3	57.0	47.6	51.5	50.3	48.6	63.3	56.4	35.5	44.4
AvgP	79.2	39.9	53.9	61.7	50.2	51.4	49.0	50.1	55.7	63.0	42.7	41.8
RandM	32.3	22.7	39.2	30.3	45.8	50.5	48.1	48.2	49.7	48.0	29.1	28.7
Rand	63.7	46.3	56.0	59.4	48.9	52.1	49.7	49.1	60.6	68.0	38.0	38.6
RandNM	87.1	75.5	71.8	81.2	50.5	53.2	50.4	50.0	66.5	73.7	49.1	48.0
TopPref	88.9	88.5	79.7	64.4	55.7	54.5	52.3	52.7	79.9	80.2	55.3	62.1
EM-MIA	99.8	97.4	98.3	99.8	47.2	50.2	51.4	50.9	88.3	80.8	88.4	77.1

Table 2: AUC-ROC results on OLMoMIA benchmark. The second block (grey) is ReCaLL-based baselines. *RandM*, *RandNM*, ReCaLL, and *TopPref* use labels in the test dataset, so comparing them with others is unfair. We report their scores for reference.

highlights ReCaLL's reliance on the availability of given non-members. Importantly, *Rand*, which uses no test labels, performs worse than Min-K%++ on average, indicating that ReCaLL alone is insufficient under a fully unsupervised setting.

RandNM is similar to the original ReCaLL (Xie et al., 2024) in most cases except for the OPT-66B model and LLaMA models with sequence length 128, probably because $n = 12$ is not optimal for these cases. *TopPref* consistently outperforms *RandNM*, demonstrating that prefix quality varies and that random prefix selection is suboptimal. This opens the door to prefix optimization (Shin et al., 2020; Deng et al., 2022; Guo et al., 2023), though finding high-quality prefixes without supervision remains challenging. Our method approximates prefix quality without labels and uses it to improve membership prediction.

6.2 OLMoMIA

Table 2 and Table 4 show results on OLMoMIA, using AUC-ROC and TPR@1%FPR as evaluation metrics respectively. EM-MIA performs nearly perfectly on *Easy* and *Medium*, similar to its performance on WikiMIA. We did not observe consistent differences across checkpoints, despite the expectation that earlier training data would be harder to detect. Therefore, we report averages across four OLMo checkpoints. In contrast, it performs close to random guessing on *Hard* and *Random* similar to MIMIR, where member and non-member distributions heavily overlap and all methods are not sufficiently better than random guessing. On *Mix-1* and *Mix-2*, EM-MIA achieves reasonable scores,

though not as high as in easier settings. In all but the hardest scenarios, EM-MIA significantly outperforms all baselines.

None of the baselines without ReCaLL-based approaches are successful in all settings, which implies that OLMoMIA is a challenging benchmark. The relative order between ReCaLL-based baselines is again consistent: *RandM* < *Avg*, *AvgP*, *Rand* < *RandNM* < *TopPref*, although none of the fully unsupervised variants are successful overall.

Interestingly, *RandNM* works reasonably well on *Mix-1* but does not work well on *Mix-2*. This is likely because non-members from *Mix-1* are from the same cluster while non-members from *Mix-1* are randomly sampled from the entire distribution. *TopPref* again outperforms *RandNM*, reinforcing that not all non-members are equally effective as prompts.

Evaluating MIA for LLMs is difficult due to unknown test-time data distributions. Benchmarks like OLMoMIA that simulate varied scenarios offer a more comprehensive lens than fixed-split benchmarks. We encourage future work to assess methods across multiple difficulty levels. While OLMoMIA is not intended as a strictly more realistic benchmark, it captures plausible conditions not reflected in prior datasets. Our results show that EM-MIA maintains strong performance across a wide spectrum of distributional overlap.

6.3 Ablation Study on Initializations and Scoring Functions

Figure 3 shows the ablation study on initialization methods (Loss, Ref, Zlib, Min-K%, Min-K%++) and prefix scoring functions (AUC-ROC, RankDist, and Kendall-Tau), using WikiMIA with length 128 and Pythia-6.9B. Each curve indicates the change of AUC-ROC calculated from the estimates of membership scores at each iteration during the expectation-maximization algorithm. In most combinations, EM-MIA converges to a similar accuracy within 4–5 iterations. In this figure, there is only one case in which AUC-ROC decreases quickly and reaches a value close to 0. It is difficult to know when this happens, but it predicts members and non-members oppositely, meaning that using negative membership scores gives a good AUC-ROC.

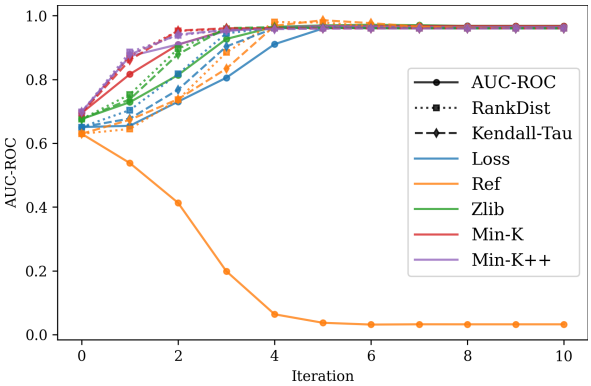


Figure 3: Performance of EM-MIA for each iteration with varying baselines for initialization and scoring functions S on WikiMIA (Shi et al., 2023) (length 128) using Pythia-6.9B (Biderman et al., 2023).

7 Conclusion

We propose EM-MIA, a membership inference method for large language models that jointly estimates membership scores and prompt effectiveness through an expectation-maximization procedure. Unlike prior work that relies on labeled non-members or assumes prompt quality in advance, EM-MIA operates in a fully unsupervised gray-box setting, making it suitable for more realistic deployment scenarios. Our method outperforms ReCaLL, even without its strong assumptions, and achieves state-of-the-art results on WikiMIA. EM-MIA is modular and flexible, allowing different initialization strategies, scoring rules, and convergence criteria depending on the application context.

To support more rigorous and controlled evaluation, we introduce OLMoMIA, a new benchmark built from the OLMo pretraining pipeline that allows fine-grained control over distributional overlap between members and non-members. Through comprehensive experiments, we show that EM-MIA is robust across a wide range of difficulty settings, while also identifying scenarios where all existing methods struggle, particularly when member and non-member distributions are nearly identical. Our findings highlight the importance of evaluating MIA methods under diverse and ambiguous conditions, and suggest that future progress will require methods that adapt to both prompt variability and fine-grained data overlap.

Limitations

Our paper focuses on detecting LLMs' pre-training data with the gray-box access where computing the probability of a text from output logits is possible. However, many proprietary LLMs are usually further fine-tuned (Ouyang et al., 2022; Chung et al., 2024), and they only provide generation outputs, which is the black-box setting. We left the extension of our approach to MIAs for fine-tuned LLMs (Song and Shmatikov, 2019; Jagannatha et al., 2021; Mahloujifar et al., 2021; Shejwalkar et al., 2021; Mireshghallah et al., 2022; Tu et al., 2024; Feng et al., 2024) or LLMs with black-box access (Dong et al., 2024; Zhou et al., 2024; Kaneko et al., 2024) as future work.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

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.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Kent K Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to chatgpt/gpt-4. *arXiv preprint arXiv:2305.00118*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Debeshee Das, Jie Zhang, and Florian Tramèr. 2024. Blind baselines beat membership inference attacks for foundation models. *arXiv preprint arXiv:2406.16201*.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv:2402.15938*.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*.

Cédric Eichler, Nathan Champeil, Nicolas Anciaux, Alexandra Bensamoun, Heber Hwang Arcolezi, and José Maria De Fuentes. 2024. Nob-mias: Non-biased membership inference attacks assessment on large language models with ex-post dataset construction. *arXiv preprint arXiv:2408.05968*.

Qizhang Feng, Siva Rajesh Kasa, Hyokun Yun, Choon Hui Teo, and Sravan Babu Bodapati. 2024. Exposing privacy gaps: Membership inference attack on preference data for llm alignment. *arXiv preprint arXiv:2407.06443*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Connecting large language models

with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. 2023. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79.

Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Membership inference attack susceptibility of clinical language models. *arXiv preprint arXiv:2104.08305*.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A Choquette-Choo, and Zheng Xu. 2023. User inference attacks on large language models. *arXiv preprint arXiv:2310.09266*.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.

Masahiro Kaneko, Youmi Ma, Yuki Wata, and Naoaki Okazaki. 2024. Sampling-based pseudo-likelihood for membership inference attacks. *arXiv preprint arXiv:2404.11262*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

California State Legislature. 2018. California consumer privacy act (ccpa). https://oag.ca.gov/privacy/ccpa.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*.

Saeed Mahloujifar, Huseyin A Inan, Melissa Chase, Esha Ghosh, and Marcello Hasegawa. 2021. Membership inference on word embedding and beyond. *arXiv preprint arXiv:2106.11384*.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024. Llm dataset inference: Did you train on my dataset? *arXiv preprint arXiv:2406.06443*.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. 2024a. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2369–2385.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. 2024b. Inherent challenges of post-hoc membership inference for large language models. *arXiv preprint arXiv:2406.17975*.

Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre de Montjoye. 2024c. Copyright traps for large language models. *arXiv preprint arXiv:2402.09363*.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. *arXiv preprint arXiv:2203.03929*.

Hamid Mozaffari and Virendra J Marathe. 2024. Semantic membership inference attack against large language models. *arXiv preprint arXiv:2406.10218*.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2024. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*.

Oscar Sainz, Iker García-Ferrero, Alon Jacovi, Jon Ander Campos, Yanai Elazar, Eneko Agirre, Yoav Goldberg, Wei-Lin Chen, Jenny Chim, Leshem Choshen, et al. 2024. Data contamination report from the 2024 conda shared task. *arXiv preprint arXiv:2407.21530*.

Virat Shejwalkar, Huseyin A Inan, Amir Houmansadr, and Robert Sim. 2021. Membership inference attacks against nlp classification models. In *NeurIPS 2021 Workshop Privacy in Machine Learning*.

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. *arXiv preprint arXiv:2310.16789*.

Igor Shilov, Matthieu Meeus, and Yves-Alexandre de Montjoye. 2024. Mosaic memory: Fuzzy duplication in copyright traps for large language models. *arXiv preprint arXiv:2405.15523*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

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.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206.

Charles Spearman. 1961. The proof and measurement of association between two things. *The American Journal of Psychology*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2024. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jonathan Tow. 2023. Stablelm alpha v2 models.

Shangqing Tu, Kejian Zhu, Yushi Bai, Zijun Yao, Lei Hou, and Juanzi Li. 2024. Dice: Detecting in-distribution contamination in llm's fine-tuning phase for math reasoning. *arXiv preprint arXiv:2406.04197*.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

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. *arXiv preprint arXiv:2406.15968*.

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.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. 2024. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

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.

A Using External Data

We may extend the test dataset $\mathcal{D}_{\text{test}}$ by utilizing external data to provide additional signals. Suppose we have a dataset of known members ($\mathcal{D}_{\text{m}}$), a dataset of known non-members ($\mathcal{D}_{\text{nm}}$), and a dataset of instances without any membership information ($\mathcal{D}_{\text{unk}}$). For example, $\mathcal{D}_{\text{m}}$ could be old Wikipedia documents, sharing the common assumption that LLMs are usually trained with Wikipedia. As discussed above, we target the case of $\mathcal{D}_{\text{nm}} = \phi$, or at least $\mathcal{D}_{\text{nm}} \cap \mathcal{D}_{\text{test}} = \phi$. However, we can construct it with completely unnatural texts (e.g., "*b9qx84;5zln"). $\mathcal{D}_{\text{unk}}$ is desirably drawn from the same distribution of $\mathcal{D}_{\text{test}}$ but could be from any corpus when we do not know the test dataset distribution. Finally, we can incorporate all available data for better prediction of membership scores and prefix scores: $\mathcal{D}_{\text{test}} \leftarrow \mathcal{D}_{\text{test}} \cup \mathcal{D}_{m} \cup \mathcal{D}_{\text{nm}} \cup \mathcal{D}_{\text{unk}}$.

B Comparison with ReCaLL

As explained in §3.2, the original ReCaLL (Xie et al., 2024) uses labeled data from the test dataset, which is unfair to compare with the above baselines and ours. More precisely, p_i in the prefix $p = p_1 \oplus p_2 \oplus \cdots \oplus p_n$ are known non-members from the test set $\mathcal{D}_{\text{test}}$, and they are excluded from the test dataset for evaluation, i.e., $\mathcal{D}_{\text{test}}{}' = \mathcal{D}_{\text{test}} \setminus \{p_1, p_2, \cdots, p_n\}$. However, we measure the performance of ReCaLL with different prefix selection methods to understand how ReCaLL is sensitive to the prefix choice and use it as a reference instead of a direct fair comparison.

Since changing the test dataset every time for different prefixes does not make sense and makes the comparison even more complicated, we keep them in the test dataset. A language model tends to repeat, so $\text{LL}(p_i|p; \mathcal{M}) \simeq 0$. Because $\text{LL}(p_i|p; \mathcal{M}) \ll 0$, $\text{ReCaLL}_p(p_i; \mathcal{M}) \simeq 0$. It is likely to $\text{ReCaLL}_p(p_i; \mathcal{M}) \ll \text{ReCaLL}_p(x; \mathcal{M})$ for $x \in \mathcal{D}_{\text{test}} \setminus \{p_1, p_2, \cdots, p_n\}$, meaning that ReCaLL will classify p_i as a non-member. The effect would be marginal if $|\mathcal{D}_{\text{test}}| \gg n$. Otherwise, we should consider this when we read numbers in the result table.

The original ReCaLL (Xie et al., 2024) is similar to *RandNM*, except they report the best score after trying all different n values, which is again unfair. The number of shots n is an important hyperparameter determining performance. A larger n generally leads to a better MIA performance but increases computational cost with a longer p.

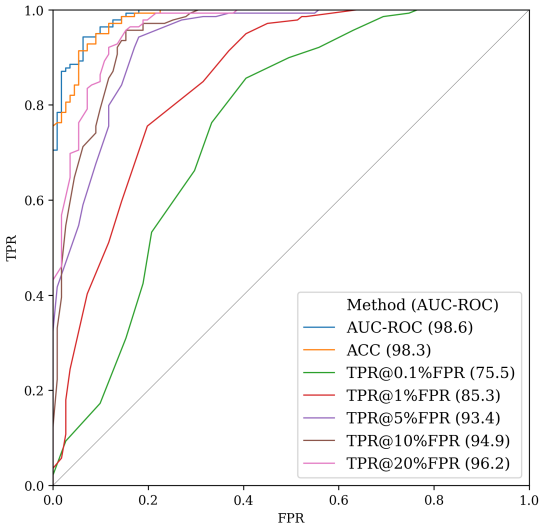


Figure 4: ROC curves for MIA using the negative prefix score as the membership score, evaluated with different metrics for prefix scores in the oracle setting on WikiMIA (Shi et al., 2023) (length 128) using Pythia-6.9B (Biderman et al., 2023).

C Metrics for Prefix Scores

Figure 4 shows ROC curves when negative prefix scores, computed using different metrics, are used directly as membership scores. We compare prefix scoring metrics including AUC-ROC, Accuracy, and TPR@k%FPR for $k \in \{0.1, 1, 5, 10, 20\}$. Among them, using AUC-ROC to compute prefix scores yields the best result, achieving 98.6% AUC-ROC for membership inference.

D Formulation of OLMoMIA Settings

After the filtering of removing close points, let member clusters as C_i^m for $i \in [1, K]$ and non-member clusters as C_j^{nm} for $j \in [1, K]$. These clusters satisfy $d(x, y) > 0.6$ for all $x, y \in C_i^m$ and $d(x, y) > 0.6$ for all $x, y \in C_j^{nm}$. The following equations formalize how we construct different settings of OLMoMIA:

- *Random*: $\mathcal{D}_{\text{random}} = \mathcal{D}^{\text{m}}_{\text{random}} \cup \mathcal{D}^{\text{nm}}_{\text{random}}$
- *Easy*: $\mathcal{D}_{\text{easy}} = \mathcal{D}^{\text{m}}_{\text{easy}} \cup \mathcal{D}^{\text{nm}}_{\text{easy}}$, where $i_{easy}, j_{easy} = \arg\max_{(i,j)} \mathbb{E}_{x \in C_i, y \in C_j} d(x, y)$, $\mathcal{D}^{\text{m}}_{\text{easy}} = \arg\text{topk}_x \mathbb{E}_{y \in C^{nm}_{j_{easy}}} d(x, y)$, and $\mathcal{D}^{\text{nm}}_{\text{easy}} = \arg\text{topk}_y \mathbb{E}_{x \in C^{m}_{i_{easy}}} d(x, y)$
- *Hard*: $\mathcal{D}_{\text{hard}} = \mathcal{D}^{\text{m}}_{\text{hard}} \cup \mathcal{D}^{\text{nm}}_{\text{hard}}$, where $i_{hard}, j_{hard} = \arg\min_{(i,j)} \mathbb{E}_{x \in C_i, y \in C_j} d(x, y)$, $\mathcal{D}^{\text{m}}_{\text{hard}} = \arg\text{topk}_x -\mathbb{E}_{y \in C^{nm}_{j_{hard}}} d(x, y)$, and

$\mathcal{D}^{\mathrm{nm}}_{\mathrm{hard}} = \arg\mathrm{topk}_y - \mathbb{E}_{x \in C^m_{i_{hard}}} d(x, y)$

- *Medium*: $\mathcal{D}_{\mathrm{medium}} = \mathcal{D}^{\mathrm{m}}_{\mathrm{medium}} \cup \mathcal{D}^{\mathrm{nm}}_{\mathrm{medium}}$, where $i_{medium}, j_{medium} = \mathrm{median}_{(i,j)}\, \mathbb{E}_{x \in C_i, y \in C_j} d(x, y)$, $\mathcal{D}^{\mathrm{m}}_{\mathrm{medium}} \subset C^m_{i_{medium}}$, and $\mathcal{D}^{\mathrm{nm}}_{\mathrm{medium}} \subset C^{nm}_{j_{medium}}$
- *Mix-1*: $\mathcal{D}_{\mathrm{mix}-1} = \mathcal{D}^{\mathrm{m}}_{\mathrm{random}} \cup \mathcal{D}^{\mathrm{nm}}_{\mathrm{hard}}$
- *Mix-2*: $\mathcal{D}_{\mathrm{mix}-2} = \mathcal{D}^{\mathrm{m}}_{\mathrm{hard}} \cup \mathcal{D}^{\mathrm{nm}}_{\mathrm{random}}$

E Computational Costs

MIAs for LLMs only do inference without any additional training, so they are usually not too expensive. Therefore, MIA accuracy is typically prioritized over computational costs as long as it is reasonably feasible. Nevertheless, maintaining MIAs' computational costs within a reasonable range is important. Computations on all our experiments with the used datasets (WikiMIA and OLMoMIA) were manageable even in an academic setting. We compare computational complexity between EM-MIA and other baselines (mainly, ReCaLL) and describe how computational costs of EM-MIA can be further reduced below.

EM-MIA is a general framework in that the update rules for prefix scores and membership scores can be designed differently (as described in §3), and they determine the trade-off between MIA accuracy and computational costs. For the design choice described in Algorithm 1 that was used in our experiments, EM-MIA requires a pairwise computation $LL_p(x)$ for all pairs (x, p) once, where $x, p \in \mathcal{D}_{\mathrm{test}}$. These values are reused to calculate the prefix scores in each iteration without recomputation. The iterative process does not require additional LLM inferences. The time complexity of EM-MIA is $O(D^2L^2)$, where $D = |\mathcal{D}_{\mathrm{test}}|$ and L is an average token length of each data on $\mathcal{D}_{\mathrm{test}}$, by assuming LLM inference cost is quadratic to the input sequence length due to the Transformer architecture. In this case, EM-MIA does not have other tuning hyperparameters, while Min-K% and Min-K%++ have K and or ReCaLL has n. This is more reasonable since validation data to tune them is not given.

Of course, the baselines other than ReCaLL (Loss, Ref, Zlib, Min-K%, and Min-K%++) only compute a log-likelihood of each target text without computing a conditional log-likelihood with a prefix, so they are the most efficient: $O(DL^2)$ time complexity. Since ReCaLL uses a long prefix consisting of n non-member data points, its time complexity is $O(D(nL)^2) = O(n^2DL^2)$. According to the ReCaLL paper, they sweep n from 1 to 12 to find the best n, which means $O((1^2 + 2^2 + \cdots + n^2)DL^2) = O(n^3DL^2)$. Also, in some cases (Figure 3 and Table 7 in their paper), they used $n = 28$ to achieve a better result. In theory, it may seem EM-MIA does not scale well with respect to D. Nevertheless, the amount of computation and time for EM-MIA with $D \sim 1000$ is not significantly larger than ReCaLL, considering the n factor.

Moreover, ReCaLL requires $O(n^2)$ times larger memory than others including EM-MIA, so it may not be feasible for hardware with a small memory. In this sense, EM-MIA is more parallelizable, and we make EM-MIA faster with batching. Lastly, there is room to improve the time complexity of our method. We have not explored this yet, but for example, we may compute ReCaLL scores on a subset of the test dataset to calculate prefix scores as an approximation of our algorithm. We left improving the efficiency of EM-MIA as future work.

F TPR@1%FPR Results

TPR@low FPR is a useful MIA evaluation metric (Carlini et al., 2022) in addition to AUC-ROC, especially when developing a new MIA and comparing it with other MIAs. Due to the space limitation in the main text, we put TPR@low FPR here: Table 3 for WikiMIA and Table 4 for OLMoMIA.

Method	Mamba-1.4B			Pythia-6.9B			LLaMA-13B			NeoX-20B			LLaMA-30B			OPT-66B			Average		
	32	64	128	32	64	128	32	64	128	32	64	128	32	64	128	32	64	128	32	64	128
Loss	4.7	2.1	1.4	6.2	2.8	3.6	4.7	4.2	7.9	10.3	3.5	4.3	4.1	5.3	7.2	6.5	3.5	3.6	6.1	3.6	4.7
Ref	0.5	0.7	0.7	1.6	1.1	1.4	2.3	3.9	2.9	3.1	2.5	1.4	1.3	2.5	3.6	1.8	1.8	0.7	1.8	2.1	1.8
Zlib	4.1	4.9	7.2	4.9	6.0	11.5	5.7	8.1	12.9	9.3	6.3	5.0	4.9	9.5	10.1	5.7	7.0	11.5	5.8	7.0	9.7
Min-K%	7.0	4.2	5.8	8.8	3.9	7.2	5.2	6.0	15.1	10.6	3.9	7.2	4.7	7.0	5.8	9.0	7.7	8.6	7.5	5.5	8.3
Min-K%++	4.1	7.0	1.4	5.9	10.6	10.1	10.3	12.0	25.2	6.2	9.5	1.4	8.3	6.7	9.4	3.6	12.0	13.7	6.4	9.6	10.2
Avg	3.9	0.4	5.0	8.0	1.1	7.9	3.1	7.0	6.5	6.2	2.1	8.6	2.8	6.7	8.6	2.6	2.1	4.3	4.4	3.2	6.8
AvgP	0.5	0.4	0.7	1.8	0.4	0.0	0.0	0.7	0.0	1.3	0.7	0.0	0.0	0.0	2.9	2.1	12.3	24.5	0.9	2.4	4.7
RandM	0.8	0.1	0.6	0.9	0.0	1.9	0.2	0.4	7.6	0.5	0.3	1.6	0.4	0.6	8.1	0.7	0.1	0.9	0.6	0.2	3.4
Rand	3.7	3.9	2.4	2.3	3.2	7.6	1.6	2.7	7.3	4.4	5.0	4.7	1.6	3.2	7.9	2.1	3.2	3.2	2.6	3.5	5.5
RandNM	19.2	8.3	15.4	12.6	10.5	18.7	18.5	17.2	7.5	12.9	11.6	12.5	13.8	18.7	8.1	5.0	5.0	6.6	13.7	11.9	11.5
TopPref	12.7	4.2	25.2	16.0	1.4	29.5	14.2	9.2	7.9	13.4	13.7	20.9	27.1	29.9	8.6	3.9	5.6	9.4	14.6	10.7	16.9
Xie et al. (2024)	11.2	11.0	4.0	28.5	20.7	33.3	13.3	30.1	26.3	25.3	6.9	30.3	18.4	18.3	1.0	8.3	5.3	6.1	17.5	15.4	16.9
EM-MIA	54.0	47.9	51.8	50.4	56.0	47.5	66.4	75.7	58.3	51.4	64.1	59.0	61.5	66.2	71.9	83.5	73.2	39.6	61.2	63.8	54.7

Table 3: TPR@1%FPR results on WikiMIA benchmark. The second block (grey) is ReCaLL-based baselines. *RandM*, *RandNM*, ReCaLL, and *TopPref* use labels in the test dataset, so comparing them with others is unfair. We report their scores for reference. We borrow the original ReCaLL results from Xie et al. (2024) which is also unfair to be compared with ours and other baselines.

Method	Easy		Medium		Hard		Random		Mix-1		Mix-2	
	64	128	64	128	64	128	64	128	64	128	64	128
Loss	2.8	12.8	7.2	1.4	0.1	1.2	1.3	0.7	7.2	1.7	0.0	0.7
Ref	6.2	4.0	4.9	0.6	1.0	0.9	1.2	1.2	8.4	0.5	0.2	1.6
Zlib	2.0	9.8	6.7	1.1	0.2	1.6	0.9	0.7	6.4	1.7	0.0	0.7
Min-K%	1.3	6.5	5.8	1.4	0.1	1.3	1.1	0.7	6.1	2.0	0.0	0.7
Min-K%++	1.4	8.0	5.0	0.7	0.4	1.0	1.0	0.4	5.0	0.9	0.0	0.5
Avg	4.1	11.5	4.0	1.7	0.2	2.2	1.2	0.6	6.1	2.2	0.0	0.9
AvgP	11.7	0.1	2.6	7.2	0.7	1.6	0.7	1.4	4.8	12.1	0.1	0.0
RandM	3.0	4.9	2.4	1.1	0.4	2.2	0.9	0.8	7.6	1.3	0.0	0.4
Rand	4.3	7.8	3.7	1.7	0.4	2.7	1.0	0.8	10.6	3.0	0.0	0.7
RandNM	16.9	14.2	5.2	1.8	0.3	1.9	1.0	0.8	9.2	2.9	0.0	1.1
TopPref	22.0	16.6	6.3	1.9	0.4	2.2	1.1	1.4	8.1	5.1	0.0	0.5
EM-MIA	95.0	52.1	79.8	96.7	1.8	1.0	1.1	1.4	12.2	3.8	14.8	4.3

Table 4: TPR@1%FPR results on OLMoMIA benchmark. The second block (grey) is ReCaLL-based baselines. *RandM*, *RandNM*, ReCaLL, and *TopPref* use labels in the test dataset, so comparing them with others is unfair. We report their scores for reference.