Detecting Training Data of Large Language Models via Expectation Maximization

Gyuwan Kim[1*] Yang Li[2] Evangelia Spiliopoulou[2] Jie Ma[2] William Yang Wang[1*]
[1]University of California, Santa Barbara [2]AWS AI Lab
*Work done while at AWS AI Labs

EACL 2026 RABAT • MOROCCO
Mars • March 24-29, 2026 • مارس

Membership Inference Attacks for LLMs

- Motivation
  - LLMs shows remarkable performance and widely deployed, but raise concerns on reliability, fairness, privacy, safety.
  - Training data drives LLMs' behavior but the exact composition is often undisclosed. Training data prevalently includes undesirable data such as test sets, proprietary contents, or personal information.
- MIA aims to **identify whether a specific data point has been seen while training a target model**: member or non-member.

Target model ⇄ Test data
Member or non-member

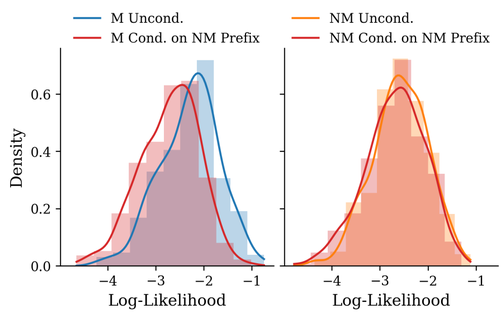
- Applications
  - Detecting data contamination for reliable evaluation
  - Auditing copyright infringement and privacy leakage
- Challenges
  - Vast amount of training data, so each instance is used only once or few times.
  - Inherent ambiguity
    - Texts are often repeated and partially overlap each other with minor difference even after decontamination and deduplication
    - Semantically similar paraphrases
  - Infeasibility of training LLMs with different datasets

Evaluation of MIA for LLMs

- Evaluation metrics: *AUC-ROC* and *TPR@low FPR*
- Common benchmarks
  - **WikiMIA**: Wikipedia event documents before (member) and after (non-member) the time cutoff based on the model release data. Some papers argue existing MIAs can achieve good performance on datasets with temporal shifts without indeed doing MIA.
  - **MIMIR**: Training (member) and test (non-member) split of PILE dataset, which are randomly separated from the same distribution. MIMIR is more challenging than WikiMIA and all existing methods are close to random guessing

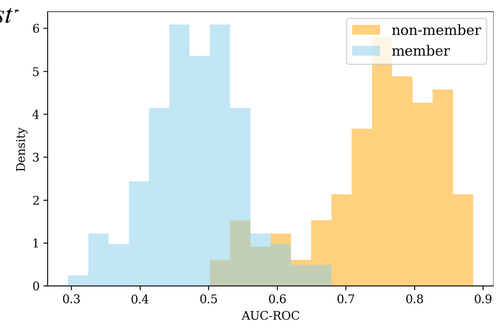
MIA Baselines (with gray-box access)

- Loss: average log-likehood (LL) of tokens, $\mathrm{LL}(x; \mathscr{M})$
- Ref: difficulty calibration using a reference model, $\mathrm{LL}(x; \mathscr{M}) - \mathrm{LL}(x; \mathscr{M}_{ref})$
- Min-K%: average LL of the top-k% tokens with the minimum token probability
- Min-K%++: token-level normalized version of Min-K%++
- **ReCaLL**: relative conditional LL with a non-member prefix, $\dfrac{\mathrm{LL}(x \mid p; \mathscr{M})}{\mathrm{LL}(x; \mathscr{M})}$
  - ReCaLL uses the ratio of the log-likelihood of a target data with and without conditioning on a *non-member prefix* (context) as a membership score, based on *empirical observation without theoretical explanation*.
  - A prefix p is a concatenation of non-members p_i ($p = p_1 \oplus \cdots \oplus p_n$), *randomly selected* from a *test set*. It assumes that (1) the ground truth non-members are available and (2) all of them are equally effective as a prefix.



Sensitivity to Prefix Choice

- We define a *prefix score* $r(p)$ as the effectiveness of p as a prefix for ReCaLL in discriminating memberships. We measure a prefix scores by how ReCaLL_p aligns well with the current membership scores f on D_{test}.
- Member and non-member distributions of prefix scores (measured in AUC-ROC with ground truth labels by using each data point as a standalone prefix) are clearly separated, suggesting that negative prefix scores can serve as effective membership scores.
- However, the true labels (i.e., the targets to be predicted) remain unknown.



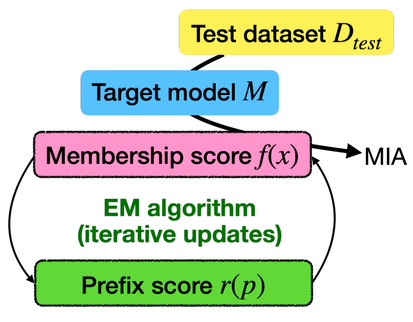
EM-MIA Framework

- Based on duality (better membership scores ↔ better prefix scores), we refine membership scores and prefix scores iteratively via an **Expectation-Maximization** algorithm until convergence.

Algorithm 1 EM-MIA

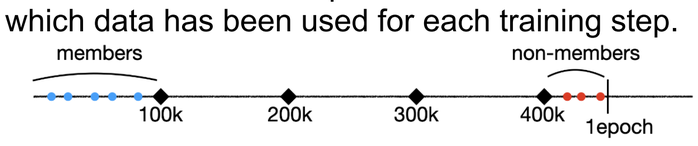
Input: Target LLM $\mathcal{M}$, Test dataset $\mathcal{D}_{\text{test}}$
Output: Membership scores $f(x)$ for $x \in \mathcal{D}_{\text{test}}$

1: Initialize $f(x)$ with an existing off-the-shelf MIA method
2: **repeat**
3: Update prefix scores $r(p) = S(\mathrm{ReCaLL}_p, f, \mathcal{D}_{\text{test}})$ 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)



OLMoMIA Benchmark

- EM-MIA works well and even performs almost perfectly on some benchmarks such as WikiMIA, while it does not work well on other datasets such as MIMIR.
- To better understand why this is the case and what are the conditions for success, we develop a new benchmark using OLMo, a series of fully open language models pre-trained with Dolma dataset. OLMo provides intermediate model checkpoints and an index to get which data has been used for each training step.



- Dataset sampling with varying difficulty
  - K-means clustering for sampled (with enough number of) members and non-members separately after embedding them.
  - *Easy*: the farthest m-nm cluster pair / instances farthest from the opposite cluster
  - *Medium*: a m-nm cluster pair with a median distance / random sampled instances
  - H*ard*: the closest m-nm cluster pair / instances closest from the opposite cluster
  - *Random*: random sampling without clustering
  - *Mix-1*: m from Random & nm from Hard
  - *Mix-2*: m from Hard and nm from Random

Experimental Setup

- ReCaLL-based baselines
  - *Avg*: $\dfrac{1}{|D_{test}|}\sum_{p \in D_{test}} ReCaLL_p(x)$, *AvgP*: $\dfrac{1}{|D_{test}|}\sum_{x \in D_{test}} ReCaLL_p(x)$
  - For a ReCaLL prefix $p = p_1 \oplus \cdots \oplus p_n$, the number of shots ($n$) is an important hyper-parameter. Usually, the larger the better, but we fix to $n = 12$.
  - *Rand*: p_i are randomly sampled data
  - *RandM*: p_i are random members
  - *RandNM*: p_i are random non-members (c.f., The original ReCaLL is similar to RandNM except they report the best score after trying all different n values which is unfair.)
  - *TopPref*: p_i have top prefix scores calculated by ground truth labels
- EM-MIA configurations
  - Min-K%++ as a default choice for initialization
  - AUC-ROC as a default scoring function
  - Update rule for membership scores: negative prefix scores

Experimental Results

- WikiMIA
  - EM-MIA achieves state-of-the-art performance on WikiMIA, significantly outperforming ReCaLL even without any given non-member test data.

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

- OLMoMIA
  - EM-MIA achieves almost perfect score on *Easy* and *Medium* like WikiMIA and similar to random guessing performance on *Hard* and *Random* like MIMIR.
  - EM-MIA get reasonably good and (not close to perfect) score on *Mix-1* and *Mix-2* though other baselines are not successful.

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

Conclusion

- We propose **EM-MIA**, a membership inference method for LLMs that jointly estimates membership scores and prompt effectiveness through an EM procedure.
- Unlike prior work relying on labeled non-members, EM-MIA operates in a **fully unsupervised** gray-box setting and significantly outperforms ReCaLL without its strong assumptions, achieving state-of-the-art results on WikiMIA.
- To enable more controlled evaluation, we introduce **OLMoMIA**, a benchmark derived from the OLMo pre-training pipeline with fine-grained control over distributional overlap.
- EM-MIA remains robust across diverse difficulty settings and reveals cases where all methods fail when member and non-member distributions are nearly identical, while also identifying scenarios where all existing methods struggle, underscoring the need for evaluating MIAs under realistic and ambiguous conditions.