

# EL-MIA: Quantifying Membership Inference Risks of Sensitive Entities in LLMs

Ali Satvaty<sup>1</sup>, Suzan Verberne<sup>2</sup>, Fatih Turkmen<sup>1</sup>

<sup>1</sup>University of Groningen, <sup>2</sup>Leiden University  
The Netherlands  
a.satvaty@rug.nl

## Abstract

Membership inference attacks (MIA) aim to infer whether a particular data point is part of the training dataset of a model. In this paper, we propose a new task in the context of LLM privacy: entity-level discovery of membership risk focused on sensitive information (PII, credit card numbers, etc). Existing methods for MIA can detect the presence of entire prompts or documents in the LLM training data, but they fail to capture risks at a finer granularity. We propose the “EL-MIA” framework for auditing entity-level membership risks in LLMs. We construct a benchmark dataset for the evaluation of MIA methods on this task. Using this benchmark, we conduct a systematic comparison of existing MIA techniques as well as two newly proposed methods. We provide a comprehensive analysis of the results, trying to explain the relation of the entity level MIA susceptibility with the model scale, training epochs, and other surface level factors. Our findings reveal that existing MIA methods are limited when it comes to entity-level membership inference of the sensitive attributes, while this susceptibility can be outlined with relatively straightforward methods, highlighting the need for stronger adversaries to stress test the provided threat model.

**Keywords:** LLMs, Membership Inference Attacks, Privacy

## 1. Introduction

LLMs have demonstrated remarkable capabilities across a wide range of natural language processing tasks, but their deployment raises growing privacy concerns (Zhao et al., 2025; Chang et al., 2024; Yan et al., 2024; Dong et al., 2024). A key risk stems from the memorization of personally identifiable information (PII) present in training data (Carlini et al., 2021; Lukas et al., 2023; Zhou et al., 2024; Satvaty et al., 2025a). Prior studies have shown that LLMs can regurgitate sensitive data verbatim or in paraphrased form, enabling membership inference attacks (MIAs) that reveal whether a specific record was used during training. Such leakage has serious implications for user privacy, regulatory compliance (e.g., GDPR, CCPA), and the safe use of LLMs in real-world applications.

While existing MIA methods have made significant progress in detecting whether entire documents were present in the training set, it is unclear to what extent they can capture risks at a finer granularity, specifically, entity-level membership risk for individual sensitive attributes such as names, dates of birth, addresses or phone numbers (Meeus et al., 2025). This is of practical relevance since attackers often seek to recover or confirm the presence of particular sensitive fields rather than entire text samples.

In this work, we introduce a framework for auditing entity-level membership risks in LLMs. Our approach targets the vulnerability of specific sensitive entities in LLM training, enabling more precise privacy risk assessments and supporting the development of stronger defenses. We construct

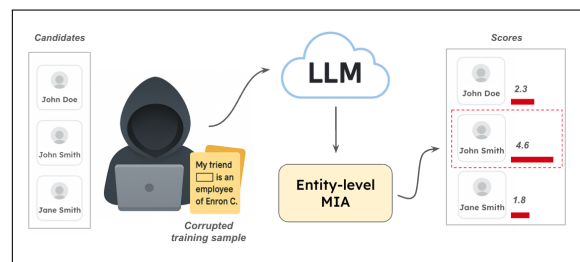


Figure 1: Entity membership inference threat model

a benchmark dataset for evaluating slot-level PII MIA, based on the existing AI4Privacy dataset (ai4Privacy, 2023). Using the newly created benchmark, we conduct a systematic comparison of existing MIA techniques, assessing their ability to detect membership of individual sensitive fields rather than long input texts. We also propose two novel methods specifically aimed at entity-level MIA attacks. Our approach reduces the noise of the token-loss signal by normalizing across a reference set and omitting the effect of prefix tokens’ loss. By empirical evaluation we compare the effectiveness of existing as well as new methods. We further perform extensive analyses regarding token length, attribute types, number of links, model size, and training dynamics.

In summary, our contributions are threefold: (1) we propose entity-level MIA tailored to the sensitive entities as a new threat model in the context of privacy in LLMs; (2) we present a new dataset, built on top of the AI4Privacy benchmark, to evalu-

ate methods on the newly proposed threat model;<sup>1</sup> (3) we evaluate the effectiveness of existing and new methods for entity-level MIA and explain the dynamics of the inference susceptibility of the samples based on various cues.

Membership unit	Work(s)	Granularity
Sequence-level	Duan et al. (2024)	~n-grams
	Carlini et al. (2021); Shi et al. (2024)	32–256 tokens
	Kaneko et al. (2024); Liu et al. (2024); Mat-tern et al. (2023), Mat-tern et al. (2023)	Variable lengths
Document-level	Meeus et al. (2024a); Shi et al. (2024); Chang et al. (2023)	Full articles, books
Dataset-level	Maini et al. (2024); Hayes et al. (2025a)	Whole benchmarks/datasets
Entity-level	<b>This work</b>	Entities + Templates

Table 1: LLM membership inference attacks by prior work, grouped by the unit of membership (sequence, document, dataset, entity). Our proposed model targets **entity-level membership**, probing whether a specific sensitive value (e.g., a name or phone number) appeared in training, given the respective context.

## 2. Background and Related Work

Membership inference was first studied in genomics, where researchers asked whether an individual’s data contributed to an aggregate dataset (Homer et al., 2008). The idea later migrated to machine learning, with the seminal attack against deep neural networks showing that models can reveal whether a specific example appeared in training (Shokri et al., 2017). Whereas traditional MIAs in machine learning typically operate on a clearly defined unit such as a single record or image, the notion of membership is more ambiguous in the text domain.

### 2.1. Granularity of text

For text data, most existing work has defined the unit of membership inference as sequences of fixed (often truncated) length. For example, Carlini et al. (2021) employ sequences of 256 tokens, whereas Shi et al. (2024) and subsequent studies (Kaneko et al., 2024; Liu et al., 2024; Mozaffari and Marathe; Wang et al., 2025; Xie et al., 2024a; Ye et al., 2024; Zhang and Wu, 2024; Zhang et al., 2024) rely on

<sup>1</sup>We release the code for creating the dataset from the original AI4Privacy data as well as our methods: <https://github.com/alistvt/el-mia-data>

random excerpts ranging between 32 and 256 tokens. When sequences are randomly injected, prior work has considered lengths of up to 80 tokens (Wei et al., 2024) or 100 tokens (Meeus et al., 2024b).

In contrast, some research aims to detect membership at a much larger scale, such as an entire book (Meeus et al., 2024a; Shi et al., 2024; Chang et al., 2023) or even a whole dataset (Maini et al., 2024; Hayes et al., 2025a). In copyright-related settings, for instance, the objective is to determine whether the full content of a source (e.g., a news article or book) was used, rather than only specific excerpts. This has led prior work (Zhang and Wu, 2024; Meeus et al., 2024a; Duarte et al., 2024) to pursue document- or dataset-level membership inference, where signals are aggregated across broader text units. This progression from fixed-length sequences to document- and dataset-level inference is further summarized in Table 1.

Yet, to the best of our knowledge, no prior work has systematically investigated finer-grained units, such as individual entities (Meeus et al., 2025). Our work aims to close this gap by introducing a benchmark for entity-level membership inference, with an emphasis on sensitive entities, as these pose the most significant privacy and security risks in practical LLM deployments.

### 2.2. Entity-level attacks on LLMs

Previous work, while not directly formulating entity-level MIAs, has examined memorization and leakage of specific entities in LLMs. Early studies introduce canary insertion and the exposure metric to quantify sequence-level memorization (Carlini et al., 2021, 2019), later adapted to NLU classifiers and auditing frameworks (Parikh et al., 2022). Similar ideas were applied to domain-specific models, e.g., probing clinical BERT for sensitive patient-condition associations and testing pseudonymization (Vakili and Dalianis, 2021, 2023).

More recently, researchers have begun to analyze entity-level memorization explicitly, showing that LLMs can reconstruct missing named entities from partial prompts (Zhou et al., 2024; Kim et al., 2023). At the same time, stronger auditing methods (Panda et al., 2025; Lukas et al., 2023; Hayes et al., 2025b) and practical PII extraction attacks (Cheng et al., 2025) demonstrate that identifiers can be surfaced.

While these approaches demonstrate valuable insights into memorization and leakage risks of sensitive entities, they fall short in comprehensively evaluating the leakage across the diverse range of sensitive entity types (e.g., names, phone numbers, account identifiers). Our work aims to fill this gap by introducing the first entity-level membership inference attack on LLMs.

### 3. Threat Model

In this section, we present our threat model and discuss how it differs from the existing threat models.

#### 3.1. Attacker’s Objective

The attacker’s objective is to determine whether a **specific entity** (e.g., a person’s full name) was part of the training data of the LLM, **given a partial or redacted sentence** (e.g., "<full name> teaches at the University of New York.") and a candidate entity value (e.g., John Doe).

Formally, let  $\mathcal{V}$  denote the set of candidate entity values and let  $\mathcal{X}$  denote the space of textual samples (strings or token sequences). A **template** is an operator  $T : \mathcal{V} \rightarrow \mathcal{X}$  that maps a value  $v \in \mathcal{V}$  to a concrete sample  $s = T(v) \in \mathcal{X}$ . Let  $D_{\text{train}} \subseteq \mathcal{X}$  be the (private) training corpus of the model. The **entity-level membership indicator** for a candidate value  $v$  under template  $T$  is the binary variable

$$M(v, T) = \begin{cases} 1 & \text{if } T(v) \in D_{\text{train}}, \\ 0 & \text{if } T(v) \notin D_{\text{train}}. \end{cases}$$

An attacker  $\mathcal{A}$  observes some side information about training samples (the template  $T$ ), and possibly additional auxiliary data or model outputs (candidate  $v$ ), and outputs either: **(1)** a score  $p \in [0, 1]$  estimating  $\Pr(M(v, T) = 1)$  or **(2)** a binary decision  $\widehat{M} \in \{0, 1\}$  (obtained by thresholding  $p$ ). The attacker’s goal is to accurately decide on  $M(v, T)$ .

#### 3.2. Attacker Capabilities

We make the following assumptions about the attacker:

**Model access.** The adversary has black-box or white-box access to the trained model; in our experiments, we assume token-level log-probabilities are queryable.

**Format knowledge.** The adversary knows the data distribution or template structure used to generate texts (e.g., sentence forms that contain PII slots).

**Candidate set.** The adversary holds a set of plausible sensitive entity values  $V$  (e.g., names, phone numbers), some of which may appear in training paired with the known templates.

**Partial sample access.** The adversary may observe a *redacted* or *canonicalized* version of a training sample, i.e., the template with the sensitive entity field masked.

#### 3.3. Applicability to Real-World Scenarios

The proposed threat model is designed to capture realistic privacy risks that arise in operational LLM

deployments. In many cases, an attacker does not have access to complete unredacted training data, but may possess partial information such as a document template or prompt structure. For example, a redacted dataset or a publicly available template (e.g., a partially anonymized medical record, legal case file, or chat log) may reveal the general format of the text while concealing the actual PII values. The attacker’s objective is then to recover the hidden fields.

In such situations, an adversary who knows the overall data format – such as the typical sequence of fields in an electronic health record, a customer support transcript, or a social media post – can construct plausible candidate values for the missing PII slots. In multi-entity contexts, an attacker may already know  $n-1$  sensitive entities appearing in related records and wish to infer the  $n$ th entity based on the correlations captured during training. For example, in a partially known patient cohort or group chat, if all but one participant’s identity is known, the model can help the attacker identify the remaining individual.

Lastly, dataset-level defenses largely rely on curation and algorithmic safeguards. A common approach is PII scrubbing, which uses Named Entity Recognition (NER) to detect and remove sensitive spans. In practice, NER systems do not achieve perfect recall across entity types, so substantial PII often remains after scrubbing (Pham et al., 2025; Keraghel et al., 2024). Moreover, Lukas et al. (2023) demonstrate that attackers can still reconstruct PII present in pre-training corpora even when models are trained on scrubbed data and, in some cases, even under differentially private training; highlighting the limits of current mitigation strategies.

#### 3.4. Distinction from Conventional Membership Inference Attacks

Our proposed threat model departs from traditional MIAs in both granularity and objective. Conventional MIAs typically operate at the level of entire samples or prompts. The attacker is given a complete input and aims to infer whether that full instance was present in the model’s training dataset. In contrast, our entity-focused threat model is designed to assess entity-level membership, targeting specific entity values (specially PIIs, e.g., names, phone numbers) inserted into redacted templates.

More specifically, rather than asking, “Was this entire sentence seen during training?”, our attack asks, “Was this particular entity value used in this specific slot in a training sample?” For example, given a sentence like “<full name> works at the Enron Corp.”, the attacker evaluates whether the model’s behavior differs when inserting “John Doe”

versus “John Smith”, and uses this information to infer whether “John Doe works at the Enron Corp.” was part of the training set, leading to violating even user privacy in case of attack success.

This distinction introduces a finer-grained and arguably more realistic privacy threat. In this work we try to highlight that models may exhibit differential behavior even when only one sensitive attribute is altered, underscoring the need for new attack methods as well as defenses that account for token-level or entity-level memorization risk.

In summary, our threat model is complementary to and more targeted than general MIAs, and is specifically motivated by the privacy risks associated with PII exposure in generative language models.

## 4. The EL-MIA Benchmark

We introduce the **EL-MIA Benchmark**, a suite comprising the *EL-MIA dataset* and the *set of trained model checkpoints*. Together, these resources provide a controlled setting for evaluating entity-level membership inference focusing on PII, enabling reproducible and systematic analysis across diverse PII types, model scales and training dynamics.

We build our dataset on top of the AI4Privacy (ai4Privacy, 2023) open-source dataset, which is designed to resemble natural, real-world text while containing a rich variety of PII. In this corpus, PII spans have already been annotated, covering multiple categories such as first names, last names, phone numbers, email addresses, and other identifying fields. These annotations allow us to directly locate and manipulate PII instances for our membership inference experiments.

Using this base data, we construct our entity-level MIA benchmark **EL-MIA** by treating each sentence as a template in which all but one PII slots remain fixed, and the remaining slot becomes the target for inference. For each target slot, we generate two sentence variants: **Member**: the original sentence containing the true PII value from the dataset, and **Non-member**: the same sentence with the target PII replaced by a randomly selected value of the same type drawn from elsewhere in the corpus. This setup ensures that the two versions differ only in the specific PII under test, enabling clean membership classification experiments. Because PII categories are varied, the dataset supports analysis across heterogeneous types, including structured identifiers (e.g., phone numbers) and free-text fields (e.g., names).

Furthermore, we release three subsets of the datasets based on the distribution of *non-member* values: **Trained**: Non-members are sampled only from the candidates on which the model has been

trained on. **Untrained**: Non-members are sampled only from the candidates on which the model has **not** been trained on. **Mix**: Non-members are sampled from the both *trained* and *untrained* sets at equal rate.

These subsets can make the analysis of the results easier to apply to the real-world scenario. The **Untrained** subset is to simulate the easiest for the attackers, because all non-member candidates have not been seen by the model, therefore the model would generally give higher perplexity to the non-members and there would be a larger gap between the member and non-members. Contrarily, **trained** subset is aimed to simulate the most difficult case, where the members are the closest to non-members. Lastly, **mix** subset aims to simulate the closest scenario to the real-world, where the attacker obtains some samples (for example using data extraction attacks) which could be memorized entities or hallucinated.

## 5. Methods and Experiments

In this section, we first benchmark prior MIA methods on the EL-MIA benchmark. After identifying their shortcomings on this task, we introduce novel methods that improve the attack success rate (ASR) over previous approaches.

### 5.1. Prior methods

We include four commonly used methods for MIA in our analysis of the EL-MIA.

**Lowest loss (Carlini et al., 2021)**. This method scores examples by the tokenized average NLL of the context: members are expected to have lower loss (higher likelihood).

**Zlib (Carlini et al., 2021)**. This method compares the model’s perplexity to the text’s zlib compression entropy, using the ratio to measure how “predictable” or memorized a sequence is, lower ratios suggest likely membership.

**Min k% probs (Shi et al., 2024)**. The method checks membership by scoring the sample token-wise, keeping the k% tokens the model is least confident about, and averaging their log-likelihood. Higher average indicates higher likelihood of membership.

**Recall (Xie et al., 2024b)**. It quantifies the change in conditional log-likelihood induced by prefixing the target example with non-member context. “Recall” method achieved the highest ASR on the various MIA benchmarks.

These methods serve as our **baselines** for evaluating EL-MIA. Since these methods were not designed for entity-level MIA, in the next section, we introduce new attacks to improve the ASR of these baselines.

## 5.2. Proposed methods

**Reference-set normalization** We propose a reference-set normalization attack method that scores how much the sensitive entity candidate  $e$  is favored by the model relative to a matched reference set of plausible alternatives ( $E_{ref} = \{e'_1, \dots, e'_N\}$ ). Given the redacted context  $S$ , we query the model in a black-box manner to obtain token-level log-likelihoods and compute a conditional score for insertion of each reference element  $e'_i$  given  $S$ . We employ *log-likelihood ratio* or its standardized variant in this decision as follows.

$$\text{LLR}(e | S) = \underbrace{\log P_\theta(e | S)}_{\text{candidate fit}} - \underbrace{\log \left( \frac{1}{N} \sum_{i=1}^N P_\theta(e'_i | S) \right)}_{\text{reference set}}$$

In our analysis, for each candidate, we build a type-consistent reference set  $E_{ref}$  (same attribute class: name, URL, phone, etc.) matched on simple surface cues with the candidate (token length, script/casing, frequency bin) while excluding the candidate itself. This design is directly practical beyond the benchmark: common entity types are readily synthesizable from public lists and format grammars (e.g., name lexicons, URL/email templates, E.164 phone patterns), enabling black-box deployment without training-data access. We present our experiments considering 5 references for each sample.

**Enhancing signal with suffix scoring** As a simple heuristic to further sharpen the membership signals, we score only the suffix immediately following the candidate span. Concretely, after the candidate boundary, we apply a continuation window of  $w$  tokens (ranging from 0 up to the full remaining number of tokens) and compute the token scores within this window, masking out the prefix. This focuses on memorized local continuations (e.g., templated phrasing) while discarding generic context that inflates noise. In all of the figures and tables we refer to this method as “Reference-set (suffix)”. We also apply this strategy to the Lowest loss baseline, which we refer to as “Lowest loss (suffix)”.

## 5.3. Experimental setup

**Models** We use the Pythia family of models (Biderman et al., 2023) of varying scales, ranging from 160M to 6.9B parameters and further pre-train them on our constructed EL-MIA dataset with the causal language modeling objective. Each model is trained for 4 epochs on our dataset, and checkpoints are stored at the end of every epoch. This design allows us to study not only the effect of model capacity on memorization and vulnerability against MIA, but also how these risks evolve

throughout the course of training. By releasing the resulting checkpoints publicly via the Hugging Face model hub, we provide the community with a standard set of baselines that can be directly used to benchmark new attack strategies or defenses on the EL-MIA dataset.<sup>2</sup> In this way, our work contributes both a carefully controlled experimental setting and a set of reproducible artifacts that we hope will accelerate research in auditing and mitigating entity-level memorization in LLMs.

**Evaluation** To evaluate the performance of different methods on the EL-MIA benchmark, we use the commonly used MIA evaluation metrics. We report the area under the attack’s ROC curve (AUC) and TPR under low FPR (5% FPR). Attack’s AUC summarizes the trade-off between true positive and false positive rates across varying decision thresholds, providing a threshold-independent indicator of separability between in-training and out-of-training samples. We report the weighted average AUC and TPR at low FPR for different PII categories present in the EL-MIA dataset to gain a better understanding of the performance of the attacks on different PII types.<sup>3</sup>

## 6. Results

In this section, we present experimental results guided by six research questions:

**RQ1.** How do the prior and our proposed methods perform on the EL-MIA? (6.1)

**RQ2.** Can the performance of the attack be explained by the length of sensitive attribute, prefix, or the whole sample? (6.2)

**RQ3.** Do different sensitive attributes show different susceptibility to EL-MIA methods? (6.3)

**RQ4.** What is the effect of the number of entities in a sample? (6.4)

**RQ5.** How does the model size affect the susceptibility? (6.5)

**RQ6.** How does the entity level susceptibility changes w.r.t the training epochs? (6.6)

### 6.1. Comparison of the methods

The main results are presented in Table 2, comparing seven EL-MIA methods across five model sizes and our three dataset subsets, after 1 epoch of training. Firstly, the proposed reference-set attacks dominate: the reference-set (suffix) variant is consistently best for every model size and subset, peaking interestingly for the 1B untrained split.

<sup>2</sup>Trained model checkpoints are accessible at <https://huggingface.co/collections/alistvt/el-mia>

<sup>3</sup>The results of global thresholding is presented at Appendix A

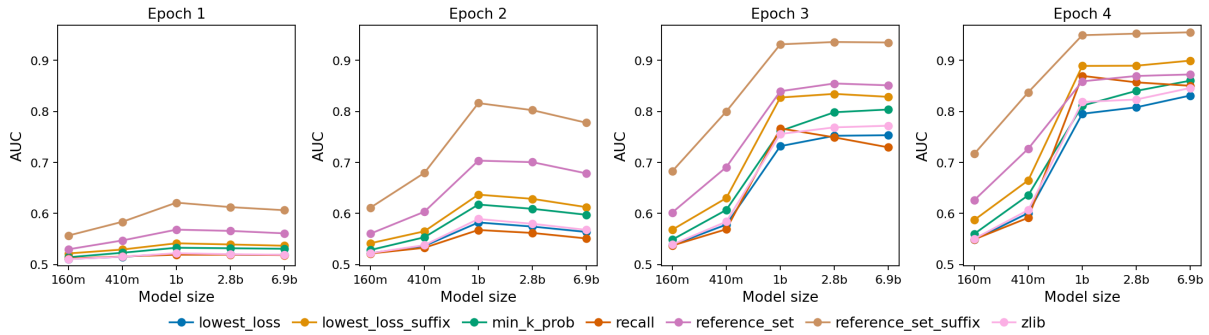


Figure 2: Analysis of all methods in terms of AUC by model size after different epochs of training

Its non-suffix version is second best, and “lowest loss (suffix)” also beats its baseline (Lowest loss). Among the baselines, Min-k% prob generally outperforms others, but still remains below the task-specific proposed methods. Across subsets, scores are typically highest on untrained, then mix, and lowest on trained as expected (see Section 4). Overall, these results highlight the limitations of the general MIA methods to the proposed threat model.

## 6.2. Effect of token count

We analyze the TPR based on the sensitive attribute token count, prefix token count (the text that appears earlier than the sensitive attribute), and the whole context token count. Correlation results for the “reference-set” attack are summarized in Table 3. We find that prefix tokens length starts contributing positively to the attack success after epoch 1 for all model sizes and larger models generally utilize it more. On the other hand, for the context length and candidate length most of the models do not show any correlations and the two largest models show to be utilizing these only at the 4th epoch.<sup>4</sup>

## 6.3. Effect of sensitive attribute types

Examining the best attack per entity type, we find that the most vulnerable categories are small, low-entropy, often single-token domains such as SEX (e.g., {male, female}), ORDINALDIRECTION ({N,E,S,W}), or CURRENCYCODE ({USD, EUR, ...}). In such cases, the model can assign sharply peaked probabilities to a handful of options, making membership cues easier to detect and leading to higher AUC/TPR.<sup>5</sup>

<sup>4</sup>Detailed statistics of each of the 56 entity types is presented in Appendix C.

<sup>5</sup>You can refer to Appendix C for detailed performance on each of the entity types.

Motivated by this, we analyze the relation between vulnerability and the number of unique candidates per attribute with a regression analysis. We find a small negative correlation ( $r = -0.150$ ) which suggests that attributes with fewer unique values tend to be more susceptible (higher AUC) to attacks, but the relation is not significant ( $p = 0.271$ ).

## 6.4. Effect of number of entities

To quantify the relation between the number entities in each sample and MIA vulnerability of the candidate of interest, we group the dataset based on the number of sensitive entities included in each sample, then we report the TPR at 5% FPR rate of each group.<sup>6</sup> Figure 3 shows the plot for the 6.9B model; the results for the other model sizes are similar. We find a negative correlation across all methods up to epoch 3, however, for large models (2.8B and 6.9B) after the third epoch the value gets positive. This suggests in the earlier epochs of training, complex samples would benefit from further privacy, however, after more training, large models would remind those samples more easily, as they would start to attend to more complex cues.

## 6.5. Model size

When we look at the susceptibility as a function of model size (at fixed epoch) and training dynamics (across epochs) (Figure 2), we observe that across epochs, AUC rises for all attacks. With the model size, AUC seems to saturate around 1B, hinting that capacity stops being the bottleneck. This could be due to the fact that larger models can memorize and stabilize entity-specific likelihoods faster, and more epochs amplify this separation. The 1B model at the first 2 epochs remains the most susceptible model which at first is counter-intuitive, however this phenomena is consistent

<sup>6</sup>Note that the same approach cannot be done for the AUC, as it cannot be measured per sample.

Model Size	Method	mix		untrained		trained	
		AUC	TPR <sub>5%</sub>	AUC	TPR <sub>5%</sub>	AUC	TPR <sub>5%</sub>
160m	Lowest loss	51.07	5.56	51.36	5.72	50.80	5.45
	Zlib	51.03	5.53	51.33	5.51	50.77	5.48
	Min k% prob	51.39	5.69	51.89	5.96	50.93	5.58
	Recall	51.13	5.78	51.14	5.59	51.14	5.91
	Lowest loss (suffix)	52.14	6.03	52.62	6.29	51.68	5.76
	Reference-set	52.93	6.88	53.84	6.99	52.10	5.84
	Reference-set (suffix)	<b>55.64</b>	<b>7.75</b>	<b>56.81</b>	<b>7.92</b>	<b>54.51</b>	<b>7.65</b>
410m	Lowest loss	51.51	5.63	51.83	5.81	51.15	5.60
	Zlib	51.51	5.75	51.86	5.84	51.16	5.75
	Min k% prob	52.29	6.19	52.90	6.38	51.66	5.84
	Recall	51.62	6.14	51.70	6.32	51.35	6.15
	Lowest loss (suffix)	52.92	6.64	53.50	6.75	52.33	6.30
	Reference-set	54.66	7.40	55.88	7.62	53.55	6.72
	Reference-set (suffix)	<b>58.36</b>	<b>8.81</b>	<b>59.60</b>	<b>8.95</b>	<b>57.10</b>	<b>8.22</b>
1b	Lowest loss	52.12	5.82	52.56	6.06	51.64	5.81
	Zlib	52.19	6.15	52.67	6.38	51.70	6.06
	Min k% prob	53.28	6.65	54.05	7.18	52.47	6.50
	Recall	51.85	6.47	52.14	6.74	51.72	6.26
	Lowest loss (suffix)	54.18	7.43	54.94	7.82	53.31	6.94
	Reference-set	56.86	7.93	58.42	9.33	55.16	7.22
	Reference-set (suffix)	<b>62.19</b>	<b>11.09</b>	<b>63.58</b>	<b>11.88</b>	<b>60.55</b>	<b>10.40</b>
2.8b	Lowest loss	51.95	6.00	52.36	6.25	51.52	5.76
	Zlib	52.00	6.14	52.43	6.38	51.56	6.02
	Min k% prob	53.16	6.85	53.88	7.23	52.46	6.47
	Recall	52.05	6.61	52.03	6.64	51.63	6.64
	Lowest loss (suffix)	53.94	7.15	54.64	7.56	53.15	6.82
	Reference-set	56.51	8.14	58.14	8.96	55.06	6.91
	Reference-set (suffix)	<b>61.18</b>	<b>10.78</b>	<b>62.71</b>	<b>11.30</b>	<b>59.80</b>	<b>10.04</b>
6.9b	Lowest loss	51.84	5.71	52.21	6.04	51.46	5.61
	Zlib	51.88	6.13	52.29	6.28	51.49	6.01
	Min k% prob	53.07	6.64	53.75	6.99	52.38	6.29
	Recall	51.74	6.53	52.18	6.82	51.55	6.40
	Lowest loss (suffix)	53.67	7.23	54.35	7.51	52.97	7.00
	Reference-set	56.10	7.52	57.26	8.76	54.90	6.77
	Reference-set (suffix)	<b>60.72</b>	<b>9.92</b>	<b>61.93</b>	<b>10.76</b>	<b>59.22</b>	<b>9.56</b>

Table 2: Performance of each attack (rows) on each dataset subset (columns), grouped by model size, for attribute type specific thresholding. Only 1 epoch results are shown. The methods marked in purple are the newly proposed methods. Boldface indicates the best result per model size; underline indicates the best result over all model sizes.

with the bias–variance balance: very large models haven’t specialized yet, while mid-size models already overfit to the data (Belkin et al., 2019).

## 6.6. Training dynamics

Susceptibility increases with additional epochs for all sizes, but the rate is size-dependent: larger models show a sharp rise between epochs 2 to 3 and then begin to level off by epoch 4, while smaller models grow more modestly and saturate lower. Taken together (with the model size behavior), these results challenge the earlier findings that bigger models exhibit more memorization (Carlini et al., 2023): while at the final epoch models do

scale in susceptibility, early epoch behavior can invert that ranking, with a 1B model peaking first.

## 7. Conclusions

In this paper we formulated a new threat model for auditing the presence of privacy sensitive information in the training data of LLMs: Entity-Level Membership inference attack (EL-MIA). We created a benchmark dataset and evaluated existing MIA methods as well as two newly proposed methods on the proposed dataset. We found that the existing MIA methods struggle with membership inference on the sensitive entities but our newly proposed methods outperforms all other methods

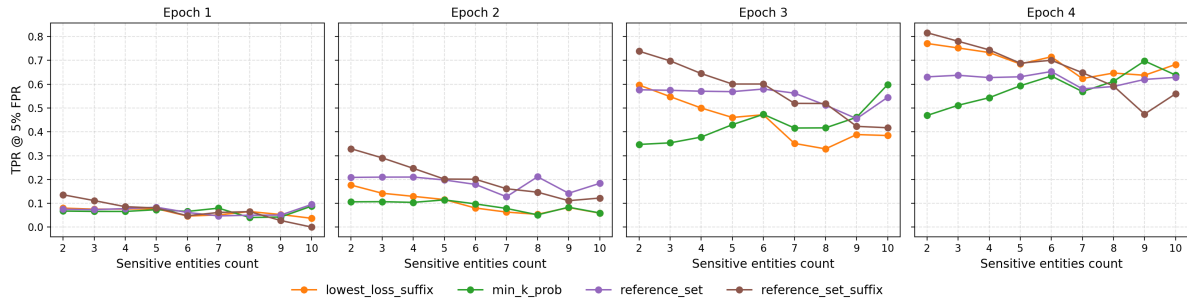


Figure 3: Attack TPR at 5% FPR explained by number of sensitive entities present in the sample, for the 6.9b model across different epochs. For brevity only best methods are shown.

Model	Epoch	Tokens length		
		Prefix	Context	Candidate
160m	1	0.17	0.15	0.17
	2	0.28	0.17	0.19
	3	0.40	0.18	0.23
	4	0.43	0.14	0.18
410m	1	0.14	-0.15	-0.11
	2	0.29	0.15	0.17
	3	0.31	0.05	0.09
	4	0.39	0.14	0.17
1b	1	0.25	-0.12	-0.12
	2	0.35	0.08	0.11
	3	0.39	0.14	0.15
	4	0.47	0.18	0.19
2.8b	1	0.14	0.02	0.10
	2	0.32	0.10	0.08
	3	0.43	0.18	0.18
	4	0.45	0.27	0.26
6.9b	1	0.14	-0.14	-0.15
	2	0.28	0.06	0.10
	3	0.47	0.21	0.20
	4	0.49	0.28	0.26

Table 3: Heat-colored Pearson’s correlations ( $r$ ) between TPR and samples’ tokens length metrics for the “reference-set” attack. Color scale maps  $r = -1$  to blue,  $r = 0$  to white,  $r = 1$  to red. Cells that are not statistically significant ( $p > 0.05$ ) are left uncolored.

in all model sizes by noticing the deviation of the model likelihood for the entity under interest to a set of crafted entities of the same type. Under properly designed attacks, the LLMs exhibit substantial vulnerability even after 1 epoch of training, underscoring both the validity of our threat model and the remaining room for improvement.

By releasing the benchmark together with resulting model checkpoints publicly via the Hugging

Face model hub, we provide the community with a standard set of baselines that can be directly used to benchmark new attack strategies or defenses on the EL-MIA dataset.<sup>7</sup> In this way, our work contributes both a carefully controlled experimental setting and a set of reproducible artifacts that we hope will accelerate research in auditing and mitigating entity-level memorization in LLMs.

Our paper is also a call for standardized evaluation of PII exposure, which is a realistic threat in operational LLM deployments. Future work should focus on expanding to other access levels and variations of the auxiliary data available to the adversary, the development of further methods, and proposing defense strategies for this threat model.

## Limitations

In this work, we study MIA arising from further *pre-training* of LLMs. Given evidence that fine-tuning can be misaligned with pre-training (Zeng et al., 2024), future work can address this gap. Our threat model assumes access to a verbatim redacted prompt (masked entity, exact template). However, sometimes the attacker may hold paraphrased or approximate templates; constructing a paraphrased EL-MIA benchmark (e.g., LLM-generated variants beyond AI4Privacy) is considered beyond the current work due to the dataset access and replicability, therefore we present the EL-MIA benchmark as a first step and the approximate MIA on the entities remains a key but challenging direction for future work. Lastly, our dataset and results are limited to English. Given cross-lingual differences in language structure and privacy dynamics (Satvaty et al., 2025b), future work should evaluate additional languages to test whether the observed patterns persist or diverge.

<sup>7</sup>Models are accessible at <https://huggingface.co/collections/alistvt/el-mia>

## Acknowledgements

This publication is part of the project LESSEN<sup>8</sup> with project number NWA.1389.20.183 of the research program NWA ORC 2020/21 which is (partly) financed by the Dutch Research Council (NWO).

We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high performance computing cluster.

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-13489.

## References

- ai4Privacy. 2023. [pii-masking-200k \(revision 1d4c0a1\)](#).
- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. [Quantifying memorization across neural language models](#). In *The Eleventh International Conference on Learning Representations*.
- Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.
- Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. [Speak, memory: An archaeology of books known to ChatGPT/GPT-4](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. [A survey on evaluation of large language models](#). *ACM Trans. Intell. Syst. Technol.*, 15(3).
- Shuai Cheng, Shu Meng, Haitao Xu, Haoran Zhang, Shuai Hao, Chuan Yue, Wenrui Ma, Meng Han, Fan Zhang, and Zhao Li. 2025. Effective {PII} extraction from {LLMs} through augmented {Few-Shot} learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 8155–8173.
- Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. 2024. [Attacks, defenses and evaluations for LLM conversation safety: A survey](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6734–6747, Mexico City, Mexico. Association for Computational Linguistics.
- Michael Duan, Anshuman Suri, Nilofar 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 *First Conference on Language Modeling*.
- André V. Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. 2024. De-cop: detecting copyrighted content in language models training data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Jamie Hayes, Iliia Shumailov, Christopher A. Choquette-Choo, Matthew Jagielski, George Kaissis, Katherine Lee, Milad Nasr, Sahra Ghalebikesabi, Nilofar Mireshghallah, Meenatchi Sundaram Mutu Selva Annamalai, Igor Shilov, Matthieu Meeus, Yves-Alexandre de Montjoye, Franziska Boenisch, Adam Dziedzic, and A. Feder Cooper. 2025a. [Strong membership inference attacks on massive datasets and \(moderately\) large language models](#).

---

<sup>8</sup><https://lessen-project.nl>

- Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, Iliia Shumailov, Milad Nasr, Christopher A. Choquette-Choo, Katherine Lee, and A. Feder Cooper. 2025b. [Measuring memorization in language models via probabilistic extraction](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9266–9291, Albuquerque, New Mexico. Association for Computational Linguistics.
- Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167.
- Masahiro Kaneko, Youmi Ma, Yuki Wata, and Naoaki Okazaki. 2024. [Sampling-based pseudo-likelihood for membership inference attacks](#).
- Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. [Recent advances in named entity recognition: A comprehensive survey and comparative study](#).
- Siwon Kim, Sangdoon Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. [ProPILE: Probing privacy leakage in large language models](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Zhenhua Liu, Tong Zhu, Chuanyuan Tan, Bing Liu, Haonan Lu, and Wenliang Chen. 2024. [Probing language models for pre-training data detection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1587, Bangkok, Thailand. Association for Computational Linguistics.
- Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE.
- Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024. [LLM dataset inference: Did you train on my dataset?](#) In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. [Membership inference attacks against language models via neighbourhood comparison](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada. Association for Computational Linguistics.
- 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, Igor Shilov, Manuel Faysse, and Yves-Alexandre De Montjoye. 2024b. Copyright traps for large language models. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. 2025. [SoK: Membership Inference Attacks on LLMs are Rushing Nowhere \(and How to Fix It\)](#). In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 385–401, Los Alamitos, CA, USA. IEEE Computer Society.
- Hamid Mozaffari and Virendra Marathe. Semantic membership inference attack against large language models. In *Neurips Safe Generative AI Workshop 2024*.
- Ashwinee Panda, Xinyu Tang, Milad Nasr, Christopher A Choquette-Choo, and Prateek Mittal. 2025. Privacy auditing of large language models. *arXiv preprint arXiv:2503.06808*.
- Rahil Parikh, Christophe Dupuy, and Rahul Gupta. 2022. [Canary extraction in natural language understanding models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 552–560, Dublin, Ireland. Association for Computational Linguistics.
- Dzung Pham, Peter Kairouz, Niloofar Mireshghallah, Eugene Bagdasarian, Chau Minh Pham, and Amir Houmansadr. 2025. [Can large language models really recognize your name?](#)
- Ali Satvaty, Suzan Verberne, and Fatih Turkmen. 2025a. [Undesirable memorization in large language models: A survey](#).
- Ali Satvaty, Anna Visman, Dan Seidel, Suzan Verberne, and Fatih Turkmen. 2025b. [Memorization is language-sensitive: Analyzing memorization and inference risks of LLMs in a multilingual setting](#). In *Proceedings of the First Workshop on*

- Large Language Model Memorization (L2M2)*, pages 106–126, Vienna, Austria. Association for Computational Linguistics.
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. [Detecting pretraining data from large language models](#). In *The Twelfth International Conference on Learning Representations*.
- 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.
- Thomas Vakili and Hercules Dalianis. 2021. Are clinical bert models privacy preserving? the difficulty of extracting patient-condition associations. In *AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021), Virtual Event, November 4-6, 2021*.
- Thomas Vakili and Hercules Dalianis. 2023. [Using membership inference attacks to evaluate privacy-preserving language modeling fails for pseudonymizing data](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 318–323, Tórshavn, Faroe Islands. University of Tartu Library.
- Cheng Wang, Yiwei Wang, Bryan Hooi, Yujun Cai, Nanyun Peng, and Kai-Wei Chang. 2025. [ConReCall: Detecting pre-training data in LLMs via contrastive decoding](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1013–1026, Abu Dhabi, UAE. Association for Computational Linguistics.
- Johnny Wei, Ryan Wang, and Robin Jia. 2024. [Proving membership in LLM pretraining data via data watermarks](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13306–13320, Bangkok, Thailand. Association for Computational Linguistics.
- Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. 2024a. [ReCaLL: Membership inference via relative conditional log-likelihoods](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8671–8689, Miami, Florida, USA. Association for Computational Linguistics.
- Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. 2024b. [ReCaLL: Membership inference via relative conditional log-likelihoods](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8671–8689, Miami, Florida, USA. Association for Computational Linguistics.
- Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. [On protecting the data privacy of large language models \(llms\): A survey](#).
- Wentao Ye, Jiaqi Hu, Liyao Li, Haobo Wang, Gang Chen, and Junbo Zhao. 2024. [Data contamination calibration for black-box LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10845–10861, Bangkok, Thailand. Association for Computational Linguistics.
- Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, Pengfei He, Yue Xing, Shuaiqiang Wang, Jiliang Tang, and Dawei Yin. 2024. [Exploring memorization in fine-tuned language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3917–3948, Bangkok, Thailand. Association for Computational Linguistics.
- Anqi Zhang and Chaofeng Wu. 2024. [Adaptive pre-training data detection for large language models via surprising tokens](#).
- Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2024. [Min-k%++: Improved baseline for detecting pre-training data from large language models](#). *arXiv preprint arXiv:2404.02936*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. [A survey of large language models](#).
- Zhenhong Zhou, Jiuyang Xiang, Chaomeng Chen, and Sen Su. 2024. [Quantifying and analyzing entity-level memorization in large language models](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'24/IAAI'24/EAAI'24*. AAAI Press.

Model Size	Method	mix		untrained		trained	
		AUC	TPR <sub>5%</sub>	AUC	TPR <sub>5%</sub>	AUC	TPR <sub>5%</sub>
160m	Lowest loss	51.30	5.26	51.58	5.28	51.08	5.26
	Zlib	51.33	5.43	51.63	5.44	51.11	5.36
	Perplexity ratio	51.19	5.46	51.42	5.55	50.98	5.44
	Min k% prob	51.56	5.50	52.07	5.55	51.15	5.53
	Lowest loss (suffix)	52.12	5.88	52.50	6.02	51.80	5.64
	Reference-set	53.08	6.09	54.38	6.28	52.23	5.82
	Reference-set (suffix)	57.52	8.53	58.90	8.48	56.18	7.68
410m	Lowest loss	51.74	5.44	52.06	5.42	51.46	5.41
	Zlib	51.80	5.60	52.15	5.62	51.50	5.53
	Perplexity ratio	52.17	5.82	52.52	6.00	51.80	5.83
	Min k% prob	52.38	5.64	52.91	5.73	51.86	5.52
	Lowest loss (suffix)	52.83	6.16	53.28	6.31	52.41	5.91
	Reference-set	54.73	6.03	56.11	6.73	53.66	6.05
	Reference-set (suffix)	60.38	9.11	61.64	10.16	58.73	8.44
1b	Lowest loss	52.30	5.49	52.74	5.58	51.90	5.36
	Zlib	52.46	5.76	52.93	5.84	52.02	5.55
	Perplexity ratio	52.69	6.28	53.22	6.54	52.15	6.20
	Min k% prob	53.22	5.90	53.93	6.08	52.52	5.80
	Lowest loss (suffix)	53.79	6.84	54.42	6.92	53.17	6.71
	Reference-set	56.47	6.78	58.11	7.40	55.24	7.12
	Reference-set (suffix)	64.16	11.70	66.44	12.99	61.54	10.30
2.8b	Lowest loss	51.94	5.79	52.31	5.88	51.57	5.54
	Zlib	52.06	5.77	52.47	5.89	51.67	5.66
	Perplexity ratio	52.34	6.18	52.83	6.11	51.85	6.14
	Min k% prob	53.28	6.28	54.03	6.49	52.60	6.15
	Lowest loss (suffix)	53.36	6.58	53.85	6.66	52.77	6.24
	Reference-set	56.46	6.75	58.06	7.55	55.04	6.71
	Reference-set (suffix)	62.10	11.05	63.91	11.74	60.40	9.92
6.9b	Lowest loss	51.85	5.57	52.18	5.75	51.52	5.37
	Zlib	51.93	5.70	52.29	5.74	51.58	5.66
	Perplexity ratio	52.18	6.38	52.60	6.48	51.76	6.18
	Min k% prob	53.17	6.16	53.85	6.22	52.51	6.02
	Lowest loss (suffix)	53.20	6.36	53.62	6.42	52.69	6.16
	Reference-set	55.89	6.88	57.31	7.30	54.70	6.57
	Reference-set (suffix)	61.58	10.30	63.02	11.34	59.74	9.01

Table 4: AUC and TPR at 5% FPR for each method (rows) and dataset subset (columns), grouped by model size, when the thresholds are selected over the whole dataset (rather than attribute type specific thresholding). Only 1 epoch results shown. The methods marked in purple are the newly proposed methods.

## A. Global thresholding

Results of the attacks when global thresholding (opposed to entity-type-specific thresholding) is used are presented in Table ???. All of the results in this table are slightly lower than their respective cell in the entity-type-specific thresholding results (Table 2).

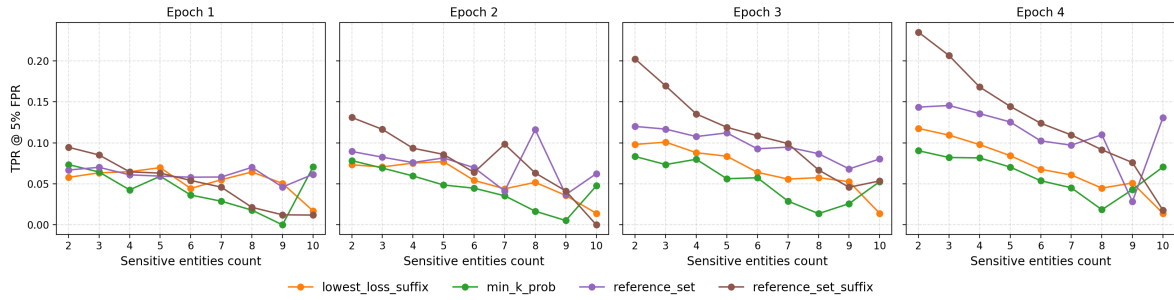


Figure 4: Attack TPR at 5% FPR explained by number of sensitive entities present in the sample, for the 160M model across different epochs. For brevity only best methods are shown.

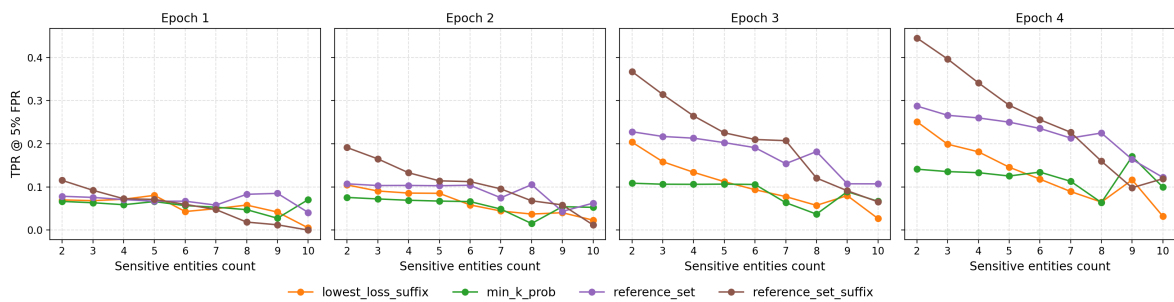


Figure 5: Attack TPR at 5% FPR explained by number of sensitive entities present in the sample, for the 410M model across different epochs. For brevity only best methods are shown.

## B. Effect of number of entities plots

Attack performance for the “mix” split, based on the number of sensitive entities in each sample is presented in Figures 4 to 8. We further process these results by calculating the Pearson’s correlation (Figure 9). As one can see in Figure 9, all methods obtain negative correlation, showing their performance degrades when more sensitive entities are present in the sentence. However, it seems that for larger models (1B onwards), and after then second epoch of training, only the Min K% Probs method is somehow utilizing the presence of more sensitive attributes in the sample. This trend also is visible in Figures 6 to 8 for the Min K% Probs method with a slightly positive slope of its curve.

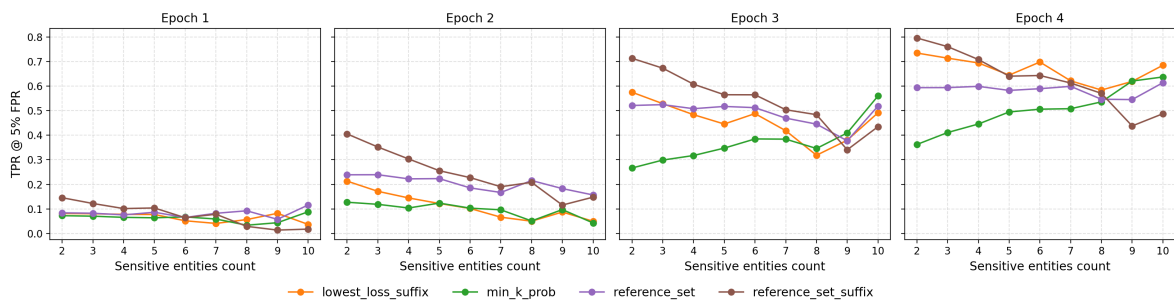


Figure 6: Attack TPR at 5% FPR explained by number of sensitive entities present in the sample, for the 1B model across different epochs. For brevity only best methods are shown.

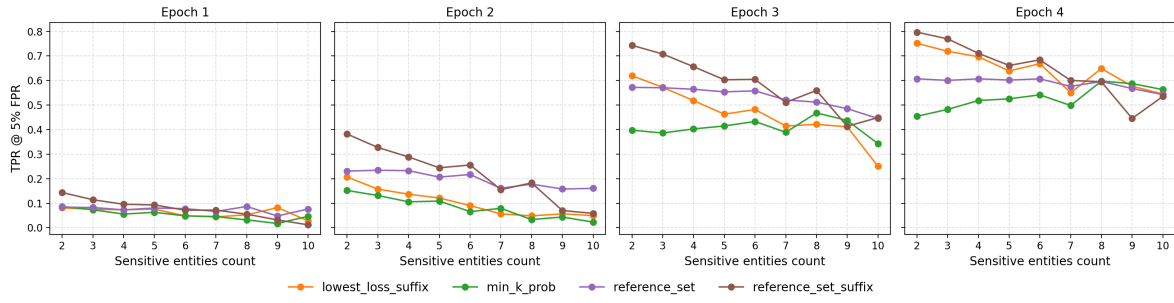


Figure 7: Attack TPR at 5% FPR explained by number of sensitive entities present in the sample, for the 2.8B model across different epochs. For brevity only best methods are shown.

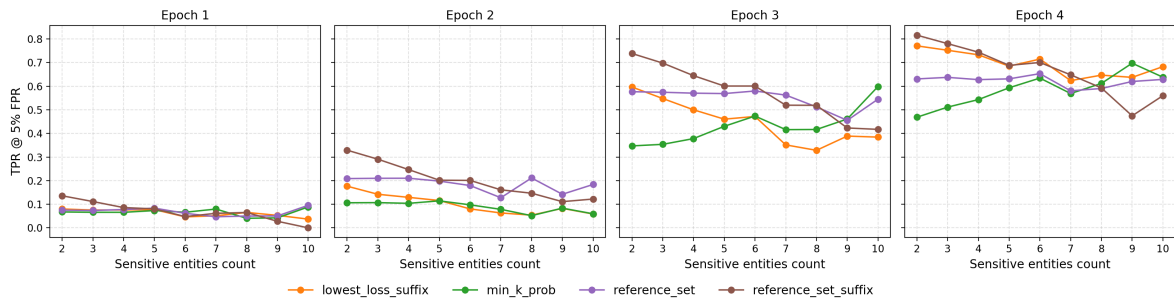


Figure 8: Attack TPR at 5% FPR explained by number of sensitive entities present in the sample, for the 6.9B model across different epochs. For brevity only best methods are shown.

### C. Attribute-specific performance

Table 5 shows the attribute-specific ASR and the statistics of different attributes in terms of their median token length.

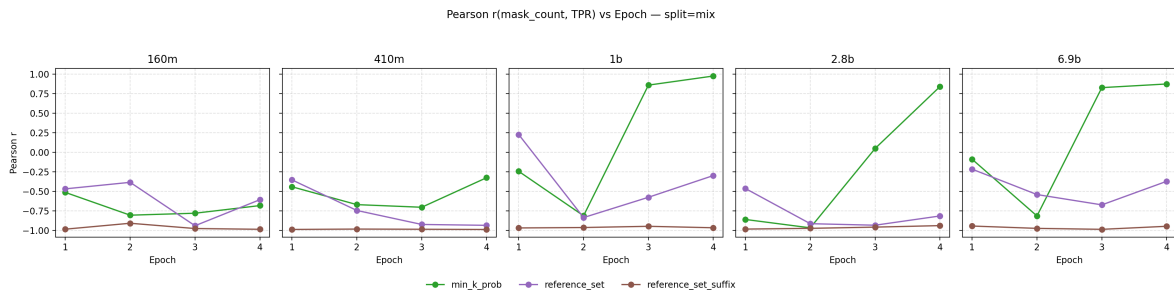


Figure 9: Pearson's correlation between attack TPR @ 5% FPR and the number of sensitive entities at each sample across different epochs and model sizes. (Derived from Figures 4 to 8)

Attribute	AUC (%)		TPR (%)		Token Len Median		
	Method	Value	Method	Value	Candidate	Prefix	Context
CREDITCARDCVV	reference_set_suffix	70.38	reference_set_suffix	21.95	1	31	39
SEX	reference_set_suffix	69.98	reference_set_suffix	14.40	1	8	37
HEIGHT	reference_set_suffix	69.88	min_k_prob	11.38	4	19	38
ORDINALDIRECTION	reference_set_suffix	69.67	reference_set_suffix	19.28	2	17	35
CURRENCYCODE	reference_set_suffix	69.03	reference_set_suffix	26.73	2	20	43
CITY	reference_set_suffix	68.71	reference_set_suffix	16.33	3	16	40
BIC	reference_set_suffix	68.05	reference_set_suffix	20.00	6	27	43
PIN	reference_set_suffix	68.03	reference_set_suffix	11.30	2	29	40
VEHICLEVRM	reference_set_suffix	68.00	reference_set_suffix	9.49	4	21	43
IPV4	reference_set	67.83	reference_set	16.92	7	23	45
URL	reference_set_suffix	67.23	reference_set_suffix	13.47	10	26	47
MIDDLENAME	reference_set_suffix	67.14	reference_set_suffix	14.48	2	5	41
FIRSTNAME	reference_set_suffix	66.52	reference_set_suffix	12.69	2	2	40
JOBAREA	reference_set_suffix	66.42	reference_set_suffix	12.53	2	15	40
PHONEIMEI	reference_set_suffix	65.91	reference_set_suffix	15.59	10	27	47
STATE	reference_set_suffix	65.78	reference_set_suffix	13.33	3	18	41
BUILDINGNUMBER	reference_set_suffix	65.30	reference_set_suffix	13.48	2	16	37
STREET	reference_set_suffix	65.14	reference_set_suffix	9.22	3	16	39
ZIPCODE	reference_set_suffix	65.12	reference_set_suffix	14.58	5	20	38
USERNAME	reference_set_suffix	64.86	reference_set	10.77	5	10	42
JOBTYPE	reference_set_suffix	64.33	reference_set_suffix	13.25	2	11	37
CURRENCYSYMBOL	reference_set_suffix	64.30	reference_set_suffix	9.62	2	19	42
VEHICLEVIN	reference_set_suffix	64.20	reference_set_suffix	12.88	10	21	44
EMAIL	reference_set_suffix	64.19	lowest_loss_suffix	10.96	10	31	46
CURRENCY	reference_set_suffix	63.42	reference_set_suffix	14.14	4	21	44
NEARBYGPSCOORDINATE	reference_set	63.11	reference_set	13.78	11	19	44
IBAN	reference_set_suffix	62.98	reference_set	13.06	11	23	46
LITECOINADDRESS	reference_set	62.96	perplexity_ratio	11.57	22	32	72
AMOUNT	reference_set_suffix	62.90	reference_set	15.69	4	19	46
ACCOUNTNUMBER	reference_set_suffix	62.33	reference_set	9.16	3	24	42
MASKEDNUMBER	reference_set_suffix	62.07	reference_set_suffix	11.46	7	21	41
COMPANYNAME	reference_set_suffix	62.02	reference_set_suffix	10.99	6	13	46
ACCOUNTNAME	reference_set_suffix	61.40	reference_set	11.76	3	18	40
EYECOLOR	reference_set_suffix	61.31	lowest_loss_suffix	12.03	2	22	38
CURRENCYNAME	reference_set_suffix	61.30	reference_set	9.48	4	21	45
LASTNAME	reference_set_suffix	61.27	reference_set_suffix	11.74	3	5	44
AGE	reference_set_suffix	60.64	lowest_loss_suffix	7.95	2	15	36
COUNTY	reference_set_suffix	60.05	reference_set_suffix	10.29	3	15	38
PHONENUMBER	reference_set_suffix	59.74	reference_set	10.14	7	34	47
BITCOINADDRESS	reference_set_suffix	59.57	min_k_prob	9.27	24	23	66
DOB	reference_set_suffix	59.41	reference_set_suffix	12.17	5	17	45
CREDITCARDNUMBER	reference_set_suffix	59.31	reference_set_suffix	11.14	7	24	40
IP	reference_set	58.57	reference_set	9.85	7	21	58
SECONDARYADDRESS	reference_set_suffix	58.54	reference_set_suffix	8.92	3	19	37
PASSWORD	reference_set	58.42	perplexity_ratio	8.85	9	33	51
GENDER	reference_set_suffix	57.99	reference_set_suffix	12.39	3	7	39
JOBTITLE	reference_set_suffix	57.75	reference_set	9.05	4	13	46
IPV6	reference_set	57.27	min_k_prob	11.38	29	26	69
DATE	reference_set_suffix	56.31	reference_set_suffix	7.96	5	20	44
SSN	reference_set_suffix	56.02	reference_set_suffix	10.00	5	21	42
TIME	reference_set	55.72	reference_set	12.53	3	23	42
PREFIX	reference_set_suffix	54.85	reference_set_suffix	7.40	2	2	44
CREDITCARDISSUER	reference_set_suffix	54.70	reference_set_suffix	11.11	2	25	40
MAC	reference_set	53.90	perplexity_ratio	7.75	14	26	52
USERAGENT	reference_set	52.43	lowest_loss_suffix	6.58	43	24	81
ETHEREUMADDRESS	perplexity_ratio	51.89	perplexity_ratio	7.08	27	32	72

Table 5: Attribute Difficulty Table of 6.9B on the Mix dataset after 1 epochs training. (ct\_med: median token length of the context. can\_med: median token length of the candidate value. pre\_med: median token length of the context before the candidate value.