# Practical Membership Inference Attacks against Fine-tuned Large Language Models via Self-prompt Calibration

Wenjie Fu
wjfu99@outlook.com
Research Center of 6G Mobile Communications, School of Cyber Science and Engineering, and Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
Wuhan, China

Huandong Wang
wanghuandong@tsinghua.edu.cn
Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University
Beijing, China

Chen Gao
chgao96@gmail.com
Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University
Beijing, China

Guanghua Liu
guanghualiu@hust.edu.cn
Research Center of 6G Mobile Communications, School of Cyber Science and Engineering, Huazhong University of Science and Technology
Wuhan, China

Yong Li
liyong07@tsinghua.edu.cn
Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University
Beijing, China

Tao Jiang
taojiang@hust.edu.cn
Research Center of 6G Mobile Communications, School of Cyber Science and Engineering, Huazhong University of Science and Technology
Wuhan, China

## ABSTRACT

Membership Inference Attacks (MIA) aim to infer whether a target data record has been utilized for model training or not. Prior attempts have quantified the privacy risks of language models (LMs) via MIAs, but there is still no consensus on whether existing MIA algorithms can cause remarkable privacy leakage on practical Large Language Models (LLMs). Existing MIAs designed for LMs can be classified into two categories: reference-free and reference-based attacks. They are both based on the hypothesis that training records consistently strike a higher probability of being sampled. Nevertheless, this hypothesis heavily relies on the overfitting of target models, which will be mitigated by multiple regularization methods and the generalization of LLMs. The reference-based attack seems to achieve promising effectiveness in LLMs, which measures a more reliable membership signal by comparing the probability discrepancy between the target model and the reference model. However, the performance of reference-based attack is highly dependent on a reference dataset that closely resembles the training dataset, which is usually inaccessible for the practical scenario. Overall, existing MIAs are unable to effectively unveil privacy leakage over practical LLMs that are fine-tuned on private datasets and overfitting-free.

To address these limitations, we propose a Membership Inference Attack based on Self-calibrated Probabilistic Variation (SPV-MIA). Specifically, recognizing that memorization in LLMs is inevitable during the training process and occurs before overfitting, we introduce a more reliable membership signal, probabilistic variation, which is based on memorization rather than overfitting. Furthermore, we introduce a self-prompt approach, which constructs the dataset to fine-tune the reference model by prompting the target LLM itself. In this manner, the adversary can collect a dataset with a similar distribution from public APIs. Extensive experiments across four representative LLMs and three datasets demonstrate that CPV-MIA can improve the attack performance in AUC by about 23.6% when compared with the best baseline.

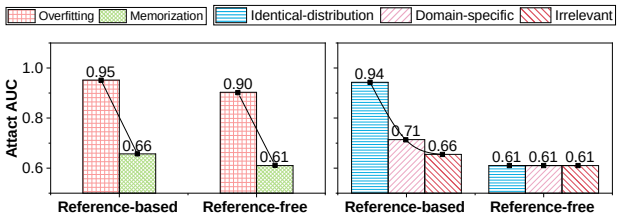
## KEYWORDS

Membership Inference Attacks; Large Language Models; Privacy and Security

## 1 INTRODUCTION

Large language models (LLMs) have been validated to have the ability to generate extensive, creative, and human-like responses when provided with suitable input prompts. Both commercial LLMs (e.g., ChatGPT [45]) and open-source LLMs (e.g., LLaMA [60]) can easily handle various complex application scenarios, including but not limited to chatbots [17], code generation [61], article co-writing [23]. Moreover, with the pretraining-finetuning paradigm gradually becoming the mainstream pipeline in the field of LLMs, small-scale organizations and even individuals can use private datasets to fine-tune over pre-trained models for downstream applications [38], which further enhances the influence of LLMs.

However, while we enjoy the revolutionary benefits raised by the popularization of LLMs, we also have to face the potential privacy risks associated with LLMs. Existing work has unveiled that the privacy leakage of LLMs can exist in almost all stages of the LLM pipeline [48]. For example, poisoning attacks can be deployed during pre-training, distillation, and fine-tuning [28, 62]. Moreover, data and model extraction attacks can be conducted through inference [7, 18]. Among these attacks, fine-tuning is widely recognized as the stage that is most susceptible to privacy leaks since the relatively small and often private datasets used for this process [69]. Therefore, this paper aims to uncover the underlying privacy concerns associated with fine-tuned LLMs through an exploration of the membership inference attack (MIA).




(a) Performance w.r.t training phase, where memorization is a stage inevitable and arises before overfitting

(b) Performance w.r.t reference dataset source, where reference datasets are utilized to fine-tune reference model for calibration

**Figure 1: Attack performances of the reference-based MIA (LiRA [39, 41]) and reference-free MIA (Neighbour Attack [36]) are unsatisfying against LLMs in practical scenarios, where LLMs are in the memorization stage and only domain-specific dataset is available. (a) Existing MIAs are unable to pose privacy leakages on LLMs that only exhibit memorization. (b) Reference-based MIA shows an exponential decrease in performance when the similarity between the reference and training datasets declines.**

MIA is an adversary model that categorizes data records into two groups: member records, which have been used in the training dataset of the target model, and nonmember records, which belong to a disjoint dataset [55]. MIAs have been well studied in classic machine learning tasks, such as classification, and reveal significant privacy risks [21]. Recently, some contemporaneous works attempt to utilize MIAs as techniques for evaluating the privacy risks of language models (LMs). For example, Mireshghallah et al. [39] first introduce a reference-based attack, Likelihood Ratio Attacks (LiRA), on Masked Language Models (MLMs), which measure the calibrated likelihood of a specific record by comparing the discrepancy on the likelihood between the target LM and the reference LM. Following this concept, Mireshghallah et al. [41] further adapt LiRA for analyzing memorization in Causal Language Models (CLMs). Moreover, Mattern et al. [36] claim that training a reference model requires access to a reference dataset that has a similar distribution as the training set of the target model, and it is almost unrealistic in practical scenarios. Therefore, they design a reference-free attack known as the Neighbour Attack to compare the discrepancy in likelihood between the target sample and its neighbour samples, which avoids the requirement of reference models. However, these methods heavily rely on several over-optimistic assumptions, including assuming the overfitting of target LLMs [36] and having access to a reference dataset from the same distribution as the training dataset [39, 41]. As a result, it remains inconclusive whether these MIAs can cause significant privacy breaches in practical scenarios.

As illustrated in Fig. 1, it respectively utilizes LiRA [41] and Neighbour Attack [36] to represent reference-based and reference-free MIAs and evaluate them from two perspectives. Firstly, as shown in Fig. 1(a), two target LLMs are fine-tuned over the same pre-trained model but stop before and after overfitting, and the reference LLMs are fine-tuned on a different dataset from the same domain. We can observe that existing MIAs cannot effectively cause privacy leaks when the LM is not overfitting. This phenomenon is addressed by the fact that the membership signal proposed by existing MIAs is highly dependent on overfitting in target LMs. They assume the member records tend to have overall higher probabilities of being sampled than non-member ones, which only satisfied overfitting models [10]. Secondly, as shown in Fig. 1(b), it validates LiRA and Neighbour Attack with three reference datasets from different sources, i.e., the dataset with the identical distribution with the member records (identical-distribution), the dataset of the same domain with the member records (domain-specific), and the dataset irrelevant to the member records (irrelevant). For the Neighbour Attack, which is a reference-free attack, the attack performance is consistently low and independent of the source of the reference dataset. For LiRA, the attack performance will exponentially decline as the similarity between the reference dataset and the target dataset declines. Thus, the reference-based MIA can not pose critical privacy leakage on LMs since a similar dataset is usually not available to the adversary model.

In this work, to address the aforementioned two limitations of existing works, we propose a Membership Inference Attack based on Self-calibrated Probabilistic Variation (SPV-MIA) composed of two according modules. First, instead of utilizing probabilities of target records as membership signals, we opt to identify member records based on memorization. Memorization is a more common phenomenon in machine learning models, which has been verified inevitable for models to arrive optimal [14]. Besides, prior work demonstrates that memorization will exist before overfitting in LLMs [59], which further improves the potential of memorization being a reliable membership signal. As existing study reveal that memorization will arise as an increased tendency in probability distribution around the member records [10], we proposed a probabilistic variation metric that can detect local maxima points via second partial derivative test [56] instantiated by a paraphrasing model. Second, although existing reference-based MIAs are challenging to reveal actual privacy risks, they demonstrate the significant potential of achieving higher privacy risks with the reference model. Therefore, we design a self-prompt approach to extract the reference dataset by prompting the target LLMs themselves and collecting the texts generated. This approach allows us to acquire the significant performance improvement brought by the reference model while ensuring the adversary model is feasible on the practical LLMs.

Overall, our contributions are summarized as follows:

- We demonstrate that detecting memorization is of great value on MIAs against overfitting-free LLMs and design a novel membership signal that detects the essential characteristics of member records memorized by LLMs by the second partial derivative test.
- We propose a self-prompt approach that collects reference datasets by prompting the target LLM with short text chunks, which will have the closely resemble distribution as the fine-tuning dataset. In this manner, the reference model fine-tuned on the reference dataset can significantly improve the attack performance without any unrealistic assumptions.
- We conducted extensive experiments to validate the effectiveness of SPV-MIA. The results suggest that SPV-MIA unveils significantly higher privacy risk across multiple fine-tuned LLMs and datasets compared with existing MIAs (about 23.6% improvement in AUC across four representative LLMs and three datasets).

## 2 PRELIMINARIES

Before delving into the technical details, we would like to introduce the Causal Language Models (CLMs) as the most representative LLMs and present a formal definition of the black-box threat model adopted in this work. Besides, the key notations used in this paper are summarized in the Appendix. A.1.

### 2.1 Causal Language Models

Since Causal Language Models (CLMs) such as GPT [50, 63], LLaMA [60] and Falcon [3] have achieved the dominant position among LLMs with various architectures, we select CLMs as representative LLMs in this work. For a given text record $\boldsymbol{x}$, it can be split into a sequence of tokens $\left[t_0, t_1, \cdots, t_{|\boldsymbol{x}|}\right]$ with variable length $|\boldsymbol{x}|$. CLM is an autoregressive language model where the model estimates the probability of the next token in a sequence given the previous tokens. Concretely, given the previous tokens $\boldsymbol{x}_{<i} = [t_0, t_1, \cdots, t_{i-1}]$, CLM aims to predict the conditional probability $p_\theta\left(t_i \mid \boldsymbol{x}_{<i}\right)$. During the training process, CLM calculates the probability of each token in a text with the previous tokens, then factorizes the joint probability of the text into the product of conditional token prediction probabilities. Therefore, the model can be optimized by minimizing the negative log probability, which can be formulated as follows:

$$\mathcal{L}_{\text{CLM}} = -\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{\left|\boldsymbol{x}^{(j)}\right|}\log p_\theta\left(t_i \mid \boldsymbol{x}_{<i}^{(j)}\right), \tag{1}$$

where $M$ denotes the number of training text records. In the process of generation, CLMs can generate coherent words by predicting one token at a time and producing a complete text using an autoregressive manner. Moreover, the pretraining-finetuning paradigm is proposed to mitigate the uncountable demands of training an LLM for a specific task [38]. This paradigm shifts to training an LLM on a shared pre-training task and then fine-tuning it to massive downstream tasks. Beiside, Multifarious parameters-efficient fine-tuning methods (e.g., LoRA [20], P-Tuning [33]) are introduced to further decrease consumption by only fine-tuning limited model parameters [11]. In this work, we focus the privacy risks in the fine-tuning phase, since the fine-tuning datasets are usually private and with smaller scales [69].

### 2.2 Threat Model

In practical applications, small companies or individuals can fine-tune public pre-trained LLMs on their private datasets for specific downstream tasks [38]. In this work, we consider an adversary who aims to infer whether a specific text record was included in the fine-tuning dataset of the target LLM. There are two mainstream scenarios investigated by previous MIA research: white-box MIA and black-box MIA. White-box MIA assumes full access to the raw copy of the target model, which means the adversary can touch and modify each part of the target model [44]. In contrast, for a black-box scenario, the adversary only approved to acquire the response results (e.g. generated texts, log probabilities) by requesting the provided service API [53], which is more realistic and aligned with practical application circumstances. Thus, we adopt the black-box scenario for evaluating existing works and our proposed method, where the adversary only receives response dictionaries from API via request texts. In more strict scenarios, only a total irrelevant dataset is available. $D$ is a dataset collected for a specific task, which can be separated into two disjoint subsets: $D_{mem}$ and $D_{non}$. The target LLM $\theta$ is fine-tuned on $D_{mem}$, and the adversary has no prior information about which data records are utilized for fine-tuning. Different from existing reference-based attacks [36, 64] require a reference from the identical distribution of training dataset, we reasonably assume that the attacker can only obtain a reference dataset $D_{refer}$ from the same domain (inferring by the task) to fine-tune the reference model. An adversary algorithm $\mathcal{A}$ is designed to infer whether a text record $\boldsymbol{x}^{(i)} \in D$ belong to the training dataset $D_{mem}$:

$$\mathcal{A}\left(\boldsymbol{x}^{(j)}, \theta\right) = \mathbb{1}\left[P\left(m^{(j)} = 1 | \boldsymbol{x}^{(j)}, \theta\right) \geq \tau\right], \tag{2}$$

where $m^{(j)} = 1$ indicates that the record $\boldsymbol{x}^{(j)} \in D_{mem}$, $\tau$ represents the threshold, and $\mathbb{1}$ denotes the indicator function.

## 3 METHODOLOGY

As demonstrated in Fig. 2, we propose a novel MIA framework against fine-tuned LLMs utilizing a calibrated probabilistic variation metric, where a paraphrasing model is introduced for assessing the probabilistic variation metric with regard to a text record, as well as a self-prompt reference model for calibrate this metric.

### 3.1 Framework

Model loss is the most widespread and straightforward metric adopted by existing MIA algorithms against machine learning model [6, 34]. As formulated in Eq. 1, the objective of an LLM is to maximize the joint probability of the text in the training set, which can also be interpreted as the negative of the loss. Therefore, some contemporaneous works employ the joint probability of the target text being sampled as the signal to evaluate the membership [2, 39, 41]. Since some records are inherently over-represented, which means even non-member records can achieve high probability in the data distribution [64]. Therefore, some works further calibrate the probability signal by comparing it with a benchmark value measured by respective methods [36, 39, 41, 64]. Overall, the **existing attack framework** can be summarized as:

$$\begin{aligned}\mathcal{A}_{exist}\left(\boldsymbol{x}, \theta\right) &= \mathbb{1}\left[\Delta p_\theta\left(\boldsymbol{x}\right) \geq \tau\right] \\ &= \mathbb{1}\left[p_\theta\left(\boldsymbol{x}\right) - \overline{p}\left(\boldsymbol{x}\right) \geq \tau\right],\end{aligned} \tag{3}$$

where $\Delta p_\theta\left(\boldsymbol{x}\right)$ is the calibrated joint probability of target text $x$, $p_\theta\left(\boldsymbol{x}\right)$ denotes the probability measured on the target model $\theta$, and $\overline{p}\left(\boldsymbol{x}\right)$ represents the benchmark probability.

However, the signal proposed by the existing attack framework is not reliable, which can be interpreted from two perspectives. First, the confidence of the probability signal is notably declined when the target model has not experienced overfitting, which guarantees the joint probabilities are higher on member texts [10]. Besides, in the fine-tuning phase of LLMs, regularization strategies are widely adopted to prevent overfitting [49, 58], and meanwhile obfuscate the probability as a metric for MIA. Second, the benchmark probability usually measured on a reference model [39, 41], which has the potential to offset the over-represented statuses of data records if the reference model can be trained on a dataset closely resemble the

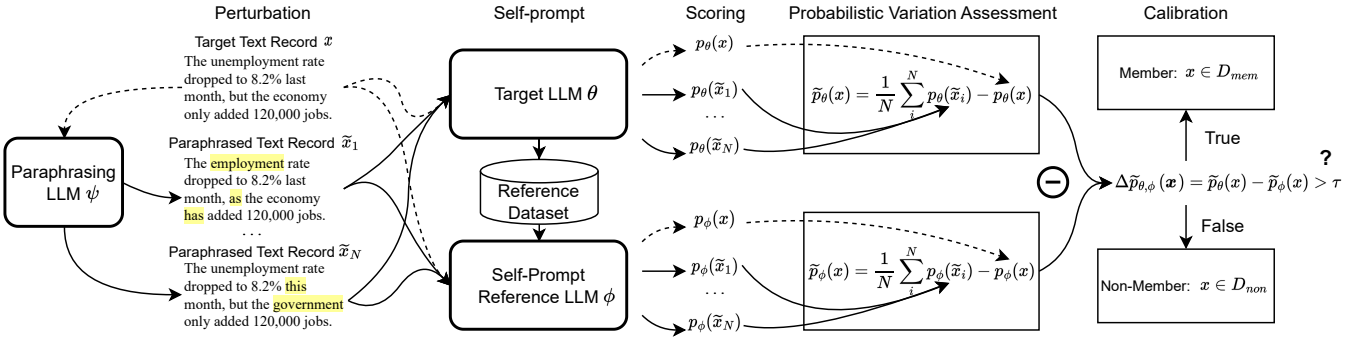



**Figure 2: The overall workflow of SPV-MIA, where includes the probabilistic variation assessment via paraphrasing model and the probabilistic variation calibration via self-prompt reference model.**

training dataset $D_{mem}$. Nevertheless, it is almost unrealistic for an adversary to obtain such a dataset, and adopting a compromising dataset will introduce noise to the benchmark probability.

In contrast to overfitting, memorization has been verified as an inevitable phenomenon for achieving optimal generalization on machine learning models [15], and it will exist before overfitting in LLMs [59]. Therefore, it will naturally be a more reliable signal for detecting member text. Memorization in generative models will cause member records to have a higher probability of being generated than neighbour records in the data distribution [10]. This principle can be shared with LLMs, as they can be considered generation models for texts. Thus, we suggest designing a more promising membership signal that can measure a value for each text record to identify whether this text is located on the local maximum in the sample distribution characterized by $\theta$. For convenience, we denote this signal as probabilistic variation $\widetilde{p}_\theta(\boldsymbol{x})$, representing the status of probabilistic variation in the local domain, and we assume the lower of $\widetilde{p}_\theta(\boldsymbol{x})$ means the more probable of the text record $\boldsymbol{x}$ locates on the local maximum. The probabilistic variation $\widetilde{p}_\theta(\boldsymbol{x})$ can be estimated as:

$$\widetilde{p}_\theta(\boldsymbol{x}) = \mathcal{F}\left(\theta, \boldsymbol{x}, \{\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}\}_{n=1}^N\right), \tag{4}$$

where $\mathcal{F}$ is the function to estimate $\widetilde{p}_\theta(\boldsymbol{x})$ and $\{\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}\}_{n=1}^N$ denotes a set of paraphrased texts of the original target text $\boldsymbol{x}$. Furthermore, we consider a reference model $\phi$ fine-tuned on a similar dataset as the target model $\theta$, which is utilized as a benchmark to calibrate the probabilistic variation measured on the target model. Formally, **our proposed attack framework** can be formulated as:

$$\begin{aligned}\mathcal{A}_{our}(\boldsymbol{x}, \theta, \phi) &= \mathbb{1}\left[\Delta\widetilde{p}_{\theta,\phi}(\boldsymbol{x}) \le \tau\right] \\ &= \mathbb{1}\left[\widetilde{p}_\theta(\boldsymbol{x}) - \widetilde{p}_\phi(\boldsymbol{x}) \le \tau\right],\end{aligned} \tag{5}$$

where $\widetilde{p}_\theta(\boldsymbol{x})$ and $\widetilde{p}_\phi(\boldsymbol{x})$ are probabilistic variations of the text record $\boldsymbol{x}$ measured on the target model $\theta$ and the reference model $\phi$ respectively. Thus, the lower value of $\Delta\widetilde{p}_{\theta,\phi}(\boldsymbol{x})$ indicates that the text record $\boldsymbol{x}$ with higher potential locates on the local maximum, and more probably drawn from the training set.

To implement our proposed adversary model into practical LLMs, we propose an attack framework with two tightly coupled modules. The workflow of our framework is depicted in Fig. 2, where we first adopt a paraphrasing model to generate paraphrased text close to the target text in the probability distribution for calculating probabilistic variations. Then, we introduce a self-prompt approach for collecting reference datasets by prompting the target LLM itself, which can be conducted without the prior knowledge of training dataset $D_{mem}$. The reference LLM will be fine-tuned on the reference dataset, which will serve as a calibrator of probabilistic variations.

## 3.2 Probabilistic Variation Assessment via Paraphrasing Model

As we discussed before, member records typically lie in the domains of local maxima of the probability function $p_\theta(\cdot)$ parameterized by LLMs. Besides, the log probability can be easily obtained from the response results of LLMs [46]. Therefore, a conceptually defined membership signal called "probability variation" is used to detect local maxima. In mathematics, the second partial derivative test is an approach in multivariable calculus commonly employed to ascertain whether a critical point of a function is a local minimum, maximum, or saddle point. In the context of our task that detects maximum points, where the hessian matrix is negative definite, i.e., all the directional second derivatives are negative. Thus, we define the probabilistic variation mentioned in Eq. 5 as the expectation of the directional second derivative:

$$\widetilde{p}_\theta(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{z}}\left(\boldsymbol{z}^\top H_p(\boldsymbol{x})\,\boldsymbol{z}\right), \tag{6}$$

where $H_p(\cdot)$ is the hessian matrix of the probability function $p_\theta(\cdot)$. Then, we further approximate the above expression with the symmetric form:

$$\boldsymbol{z}^\top H_p(\boldsymbol{x})\boldsymbol{z} \approx \frac{p_\theta(\boldsymbol{x}+h\boldsymbol{z}) + p_\theta(\boldsymbol{x}-h\boldsymbol{z}) - 2p_\theta(\boldsymbol{x})}{h^2}, \tag{7}$$

where requires $h \to 0$, and $\boldsymbol{z}$ can be interpreted as kind of "noise". Thus, $\boldsymbol{x} \pm h\boldsymbol{z}$ can be considered as adjacent text records of $\boldsymbol{x}$ in the data distribution. For simplification, we assume the distribution of "noise" $\boldsymbol{z}$ is symmetric and omit $h$, then we can reformulate Eq. 6 as follows:

$$\begin{aligned}\widetilde{p}_\theta(\boldsymbol{x}) &= \mathbb{E}_{\boldsymbol{z}}\left(p_\theta(\boldsymbol{x}+\boldsymbol{z})\right) - p_\theta(\boldsymbol{x}) \\ &= \mathbb{E}_{\widetilde{\boldsymbol{x}}\sim q(\cdot|\boldsymbol{x})}\left(p_\theta(\widetilde{\boldsymbol{x}})\right) - p_\theta(\boldsymbol{x}).\end{aligned} \tag{8}$$

where $q(\cdot \mid \boldsymbol{x})$ is a paraphrasing model that gives a distribution over $\widetilde{\boldsymbol{x}}$, slightly paraphrase the original text $\boldsymbol{x}$ and maintain the

semantics and grammar (as Eq. 7 requires $h \to 0$, which means the paraphrasing should be modest).

Based on the aforementioned discussions, and inspired by DetectGPT [42], we adopt a mask-filling model to serve as the paraphrasing model $q_\psi(\cdot \mid \boldsymbol{x})$, which is parameterized by $\psi$. Specifically, a mask-filling model such as T5 [51] aims to predict masked tokens within the input sequence. Thus, it is better at understanding context and relationships between words in a sequence, making it suitable for generating adjacent texts within the data manifold. Consequently, we randomly mask out 15% words in each target text, then employ T5-base to fill in and generate semantically coherent sentences. In this manner, we can sample a set of paraphrased texts with the of $N$. Subsequently, the probabilistic variation can be estimated as:

$$\widetilde{p}_\theta(\boldsymbol{x}) = \frac{1}{N}\sum_{n}^{N} p_\theta(\widetilde{\boldsymbol{x}}_n) - p_\theta(\boldsymbol{x}), \tag{9}$$

where $\widetilde{\boldsymbol{x}}_n \sim q_\psi(\cdot \mid \boldsymbol{x})$. Accordingly, $\widetilde{p}_\theta(\boldsymbol{x})$ is negative (i.e. lower) with higher probability for text $\boldsymbol{x}$ drawn from the training set $D_{mem}$.

## 3.3 Probabilistic Variation Calibration via Self-prompt Reference Model

Watson et al.[64] has suggested that infer the membership of a record by thresholding on a predefined metric (e.g. confidence [54], loss [68], and gradient norm [44]) will cause a high false positive rate (FPR). Since several non-member records may have high probabilities of being classified as member records simply because they are inherently over-represented in the data manifold. In other words, the metric estimated on the target model is inherently biased and has a high variance, which leads to a significant overlap in the metric distributions between members and non-members, making them more indistinguishable. To mitigate this phenomenon, Watson et al.[64] propose difficulty calibration as a general approach for extracting a much more distinguishable membership signal, which can be adapted to most metric-based MIAs by constructing their calibrated variants [39, 39, 64]. Concretely, difficulty calibration assumes an ideal reference dataset $D_{refer}$ drawn from the identical distribution as the training set $D_{mem}$ of the target model $\theta$ , and trains an ideal reference model $\phi$ with a training algorithm $\mathcal{T}$. Then, it fabricates a calibrated metric by measuring the discrepancy between metrics on the target model and reference model, and this can offset biases on membership signals caused by some over-represented records. The calibrated metric is defined as:

$$\Delta m(\boldsymbol{x}) = m_\theta(\boldsymbol{x}) - \mathbb{E}_{\phi \leftarrow \mathcal{T}(\mathcal{D}_{refer})}[m_\phi(\boldsymbol{x})], \tag{10}$$

where $\Delta m(\boldsymbol{x})$ is the calibrated version of metric, $m_\theta(\boldsymbol{x})$ and $m_\phi(\boldsymbol{x})$ are metrics measured on target and reference models, respectively.

The metric, probabilistic variation we present, is fundamentally characterized by probabilities, which can be interpreted as losses of LLMs. Consequently, the probabilistic variation is likely to be biased as well. Based on the aforementioned discussion, we naturally consider using difficulty calibration to fine-tune a reference model that acts as a metric benchmark. However, the practical deployment of this kind of attack is limited to the strong and arguably unrealistic assumption that the adversary has the approval to collect a disjoint reference dataset from the same distribution as the training dataset. Since the dataset used for fine-tuning an LLM is usually highly private and hard to extract prior knowledge. Therefore, existing work considers just utilizing the pre-trained model as the reference model [39]. In the practical scenario, at most, we can only obtain open-source datasets from the same domain or even irrelevant datasets. However, existing study have verified that adopting reference models trained on these datasets can not provide a satisfying attack performance [36].

We notice that LLMs possess revolutionary fitting and generalization capabilities, enabling them to generate a wealth of creative texts. Therefore, LLMs themselves have the potential to depict the distribution of the fine-tuning data. Thus, we consider a self-prompt approach that collects the reference dataset from the target LLM itself by prompting it with few words. Concretely, we first collect a set of text chunks with an equal length of $l$ from a public dataset from the same domain, where the domain can be easily inferred from the task of the target LLM (e.g., An LLM that serves to summary task has high probability using a summary fine-tuning dataset). Then, we utilize each text chunk of length $l$ as the prompt text and request the target LLM to generate text. All the generated text can form a dataset of size $N$, which is used to fine-tune the proposed self-prompt reference model $\phi$ over the pre-trained model. Accordingly, we can define the calibrated probabilistic variation as:

$$\Delta\widetilde{p}(\boldsymbol{x}) = \widetilde{p}_\theta(\boldsymbol{x}) - \widetilde{p}_\phi(\boldsymbol{x}), \tag{11}$$

where $\widetilde{p}_\theta(\boldsymbol{x})$ and $\widetilde{p}_\phi(\boldsymbol{x})$ are probabilistic variations measured over the target model and the self-prompt reference model.

Furthermore, in some challenging scenarios where acquiring domain-specific datasets is difficult, our self-prompt method can still effectively capture the underlying data distribution, even when using completely unrelated prompt texts. The relevant experiments will be conducted and discussed in detail in Sec. 4.3.2.

# 4 EXPERIMENTS

In this section, we evaluate the attack performance of SPV-MIA across three datasets over four LLMs, and compare it with five state-of-the-art MIAs against LLMs. The overall results validate the generic vulnerability of existing LLMs attacked by answering the following research questions:

- Dose SPV-MIA outperform the state-of-the-art MIAs?
- How does the quality of the reference model affect attack performance?
- What is the influence of different fine-tuning techniques on SPV-MIA?
- Can the existing privacy protection algorithm defend against attacks from SPV-MIA?

In addition, we also report the performance gain provided by each module via ablation study and investigate the influence of hyperparameters.

## 4.1 Experimental Settings

In this subsection, we give a brief introduction of experimental settings, including the datasets, target LLMs and baselines. The implementation details can be found in Appendix A.2.

**Table 1: AUC for detecting member texts from four LLMs across three datasets for SPV-MIA and five previously proposed methods. Bold and Underline respectively represent the best and the second-best results within each column (model-dataset pair).**

| Method | Wiki | | | | | AG News | | | | | Xsum | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GPT-2 | GPT-J | Falcon | LLaMA | **Avg.** | GPT-2 | GPT-J | Falcon | LLaMA | **Avg.** | GPT-2 | GPT-J | Falcon | LLaMA | **Avg.** |
| Loss Attack | 0.614 | 0.577 | 0.593 | 0.605 | 0.597 | 0.591 | 0.529 | 0.554 | 0.580 | 0.564 | 0.628 | 0.564 | 0.577 | 0.594 | 0.591 |
| Neighbour Attack | 0.647 | 0.612 | 0.621 | 0.627 | 0627 | 0.622 | 0.587 | 0.594 | 0.610 | 0.603 | 0.612 | 0.547 | 0.571 | 0.582 | 0.578 |
| DetectGPT | 0.623 | 0.587 | 0.603 | 0.619 | 0.608 | 0.611 | 0.579 | 0.582 | 0.603 | 0.594 | 0.603 | 0.541 | 0.563 | 0.577 | 0.571 |
| LiRA-Base | 0.710 | 0.681 | 0.694 | 0.709 | 0.699 | 0.658 | 0.634 | 0.641 | 0.657 | 0.648 | 0.776 | 0.718 | 0.734 | 0.759 | 0.747 |
| LiRA-Candidate | 0.769 | 0.726 | 0.735 | 0.748 | 0.744 | 0.717 | 0.690 | 0.708 | 0.714 | 0.707 | 0.823 | 0.772 | 0.785 | 0.809 | 0.797 |
| Our | **0.975** | **0.929** | **0.932** | **0.951** | **0.938** | **0.949** | **0.885** | **0.898** | **0.903** | **0.909** | **0.944** | **0.897** | **0.918** | **0.937** | **0.924** |

*4.1.1 Datasets.* Our experiments utilize six different datasets across multiple domains and LLM use cases, where we employ three datasets as the private datasets to fine-tune the target LLMs, and the remaining datasets as the public datasets from the exact domains. Specifically, we use the representative articles on Wikitext-103 dataset [37] to represent academic writing tasks, news topics from the AG News dataset [71] to represent news topic discussion task, and documents from the XSum dataset [43] to represent the article writing task. Besides, we utilize Wikicorpus [52], TLDR News [25], and CNNDM [19] datasets to respectively represent as the publicly accessible dataset from the same domain for each task.

*4.1.2 Target Large Language Models.* To obtain a comprehensive evaluation result, we conduct our experiments over four well-known and widely adopted LLMs as the pre-trained models with different scales from 1.5B parameters to 7B parameters:

- **GPT-2 [50]:** It is a transformer-based language model released by OpenAI in 2019, which has 1.5 billion parameters and is capable of generating high-quality text samples.
- **GPT-J [63]:** It is an open-source LLM released by EleutherAI in 2021 as a variant of GPT-3. GPT-j has 6 billion parameters and is designed to generate human-like with appropriate prompts.
- **Falcon-7B [3]:** Falcon is a family of state-of-the-art LLMs created by the Technology Innovation Institute in 2023. Falcon has 40 billion parameters, and Falcon-7B is the smaller version with less consumption.
- **LLaMA-7B [60]:** LLaMA is one of the most state-of-the-art LLM family open-sourced by Meta AI in 2023, which has outperformed other open-source LLMs on various NLP benchmarks. It has 65 billion parameters and has the potential to accomplish advanced tasks, such as code generation. In this work, we utilize the lightweight version, LLaMA-7B.

*4.1.3 Baselines.* We choose six MIAs designed for LMs to comprehensively evaluate our proposed method, including three reference-free attacks and one reference-based attack with one variant.

- **Loss Attack [68]:** A standard metric-based MIA that distinguishes member records simply by judging whether their losses are above a preset threshold.
- **Neighbour Attack [36]:** The Neighbour Attack avoids using a reference model to calibrate the loss scores and instead utilizes the average loss of plausible neighbor texts as the benchmark.
- **DetectGPT [42]:** A zero-shot machine-generated text detection method. Although DetectGPT is specially designed for LLMs-generated text detection, but has the potential to be adapted for identifying the text utilized for model training.
- **Likelihood Ratio Attack (LiRA-Base) [41]:** A reference-based attack, which adopts the pre-trained model as the reference model to calibrate the likelihood metric to infer membership.
- **LiRA-Candidate [41]:** A variant version of LiRA, which utilizes a publicly available dataset in the same domain as the training set to fine-tune the reference model.

## 4.2 Comparision with Baselines

As shown in Table. 1, we first summarize the AUC scores [5] for all baselines and SPV-MIA against four LLMs across three datasets. Furthermore, we present receiver operating characteristic (ROC) curves for SPV-MIA and the top-three best baselines on LLaMAs in Appendix A.3 for a more comprehensible presentation. Then, we can draw the following conclusions by analysing these results:

- **SPV-MIA consistently outperforms all baseline methods over all LLMs with different architectures and fine-tuning datasets:** SPV-MIA achieves the best overall attack performance with the highest average AUC of 92.4% over all scenarios. Furthermore, compared to the most competitive baseline, LiRA-Canididate, SPV-MIA has improved the AUC of the attack by 30%, even LiRA-Canididate assumes full access to the auxiliary dataset while SPV-MIA only needs some short text chunks from this dataset.
- **The overwhelming superiority of SPV-MIA compared with LiRA-Candidate demonstrates the significance of our proposed self-prompt reference model:** SPV-MIA and LiRA-Candidate both rely on the dataset from the same domain as the target LLM, which means they share the same prior information to fine-tune a reference model. However, with our proposed self-prompt approach, the reference model can learn more knowledge about the data distribution and serves as a more reliable calibrator.
- **The underwhelming attack performance of previous MIAs reveals their inability to be effectively applied to practical LLMs:** Most baseline, especially reference-free attack methods, yield a low AUC, which is only slightly better than random guesses. Furthermore, their performances on larger-scale LLMs

are worse. This phenomenon verifies the claim that existing MIAs designed for LMs can not handle LLMs with large-scale parameters.

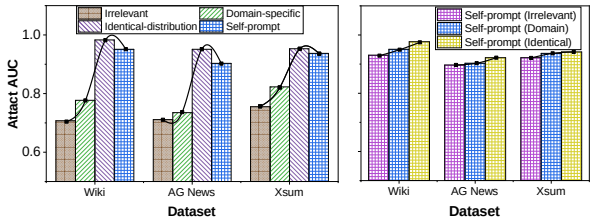
- **The privacy risk caused by MIAs on LLMs is positively correlated with the overall NLP performance of the model itself:** We found that MIAs against LLMs with similar scales like GPT-J, Falcon, and LLaMA exhibit improved attack performance as the target model performance increases. We interpret this phenomenon as follows: LLMs with stronger overall NLP performance have better learning ability, which means they are more likely to memorize records from the training set. Besides, MIAs fundamentally leverage the memorization abilities of machine learning models, making superior models more vulnerable to attacks.

## 4.3 How MIAs Rely on Reference Model Quality

In this work, a key contribution is introducing a self-prompt approach for constructing a dataset to fine-tune the reference model, which aims to improve the quality of the reference model to serve as a calibrator. Therefore, we design the experiments to verify the effectiveness of the self-prompt approach. Moreover, we investigate in detail how the quality of the reference model affects the attack performance of MIA from four aspects.

*4.3.1 Source of Reference Dataset.* In real-world scenarios, based on different prior information, adversaries can obtain datasets from different sources to fine-tune the reference model. We categorize them into three types based on their relationship with the fine-tuning dataset of the target model and sort them in ascending order of difficulty in acquisition: 1) **Irrelevant** dataset, 2) **Domain-specific** dataset, and 3) **Identical distribution** dataset. Besides, the dataset extracted by the self-prompt approach is denoted as 4) **Self-prompt** dataset. We evaluate the AUC of SPV-MIA on the aforementioned four data sources with LLaMA as the representative LLM and the results are summarized in Fig. 3(a). The experimental results indicate that the performance of the attack shows a noticeable increase along the Irrelevant, Domain, and Identical datasets. This suggests that using a reference dataset that is more similar to the target dataset can enhance the quality of the reference model. Additionally, AUC scores on self-prompt reference datasets are only marginally below Identical datasets. It verifies that our proposed self-prompt method can effectively leverage the creative generation capability of LLMs, approximate sampling text records indirectly from the distribution of the target training set.

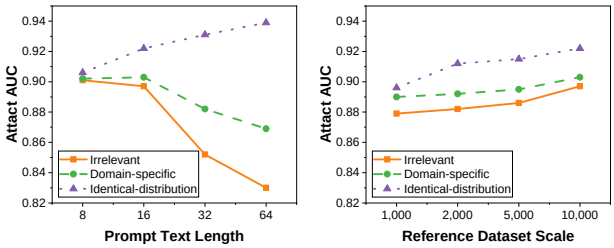
*4.3.2 Source of Self-prompt Texts.* As mentioned earlier, the self-prompt dataset is constructed by using text chunks from the domain-specific dataset to prompt target LLMs and then collecting the generated texts. These prompting text chunks are typically only a tiny fraction of the entire dataset, and thus can be collected with less cost. Compared with using domain-specific text chunks for prompting, we also evaluate the self-prompt approach with irrelevant and identical-distribution text chunks. As shown in Fig. 3(b), the AUC score increases with the relevance of the prompt texts, which aligns with the experimental results in Sec. 4.3.1. However, the self-prompt method demonstrates much lower dependence on the source of the prompt texts. We found that even when using




(a) Source of reference dataset (b) Source of self-prompt text

**Figure 3: The performances of SPV-MIA on LLaMA while utilizing different reference datasets sources and prompt texts sources, respectively.**

completely unrelated prompt texts, the performance of the attack only experiences a slight decrease. This phenomenon indicates that the self-prompt method we proposed has a high degree of versatility across adversaries with different prior information.




(a) Length of self-prompt text (b) Scale of self-prompt reference dataset

**Figure 4: The performances of SPV-MIA on LLaMA while utilizing different prompt text lengths and different scales of reference dataset.**

*4.3.3 Length of Self-prompt Texts.* In the previous experiments, we evaluated the performance of the self-prompt approach over different prompt text sources. However, in real-world scenarios, the amount of text that adversaries can obtain may vary, leading to variations in the length of the prompt texts. Therefore, we are considering a set of experiments to evaluate the attack performance across different prompt text lengths with regard to each prompt text source. The experiments are deployed over the LLaMA fine-tuned on the AG News dataset, and we set four different prompt text lengths: 8, 16, 32, and 64. The results are presented in Fig. 4, where we have discovered that texts from different sources exhibit varying trends in terms of the change in attack performance with respect to text length. Specifically, when sampling prompt texts from the identical dataset, the attack performance increases with the length of the prompt texts. When sampling from the domain dataset, the performance initially increases and then decreases with the text length. When using prompt texts sampled from an unrelated data distribution, the performance of the attack actually decreases with longer prompt texts. Therefore, we recommend setting smaller text lengths to allow LLMs to generate samples that are close to data distributions of training sets, unless adversaries can directly sample texts from the same data distribution as the training set.

**Table 2: AUC of SPV-MIA across LLaMAs fine-tuned with different fine-tuning techniques over three datasets.**

| Target Model | LoRA | Prefix Tuning | P-Tuning | $(IA)^3$ |
|---|---|---|---|---|
| # Parameters (M) | 33.55 | 5.24 | 1.15 | 0.61 |
| Wiki | 0.951 | 0.943 | 0.922 | 0.914 |
| Ag News | 0.903 | 0.897 | 0.879 | 0.873 |
| Xsum | 0.937 | 0.931 | 0.924 | 0.911 |

**Table 3: Results of Ablation Study on GPT-J and LLaMA across three datasets.**

| Target Model | Wiki | | AG News | | XSum | |
|---|---|---|---|---|---|---|
| | GPT-J | LLaMA | GPT-J | LLaMA | GPT-J | LLaMA |
| w/o PVA | 0.901 | 0.913 | 0.864 | 0.885 | 0.873 | 0.919 |
| w/o PVC | 0.648 | 0.653 | 0.632 | 0.641 | 0.653 | 0.661 |
| SPV-MIA | 0.929 | 0.951 | 0.885 | 0.903 | 0.897 | 0.937 |

*4.3.4 Scale of Self-prompt Reference Dataset.* In practical application scenarios, the public API provided by LLMs often limits the request rate to prevent malicious abuse [47]. This means that in some strict scenarios, the scale of the dataset obtained through prompting may be limited. Therefore, we investigate the performance of SPV-MIA under different scales of self-prompt reference datasets: 1,000, 2,000, 5,000, and 10,000 samples, as shown in Fig 4. Clearly, as the scale of the dataset decreases, the performance of the attack also tends to decrease to some extent. However, we have noticed that this decrease is gradual, even with only 1,000 samples, the attack performance decreases by only about 10% compared with 10,000 samples. Thus, SPV-MIA can be applied for LLMs with strict usage limitation protocols.

## 4.4 Impact of Fine-tuning Methods

As mentioned earlier, the fine-tuning algorithm in most of the experiments in this work is set to LoRA by default. However, with the pretraining-finetuning paradigm gradually gaining dominance in the field of LLMs, various Parameter-Efficient Fine-Tuning (PEFT) techniques have emerged [22]. Therefore, in order to evaluate the impact of different PEFT techniques on MIAs, we evaluate our proposed method with LLaMAs fine-tuned with different PEFT techniques. Specifically, we choose LoRA [20], Prefix Tuning [31], P-Tuning [33] and $(IA)^3$ [32] as four representative PEFT techniques, which have been widely adopted. Then, we present the number of trainable parameters as well as the AUC score of SPV-MIA across three datasets in the Table. 2. We can conclude that the risk of MIAs against LLMs is positively correlated with the number of trainable parameters during the fine-tuning process. We hypothesize that this is because as the number of trainable parameters increases, the model retains more complete memory of the training set samples, making it more vulnerable to attacks.

## 4.5 Ablation Study

In the previous experiments, we have validated the superiority of our proposed SPV-MIA over existing algorithms, as well as its versatility in addressing various challenging scenarios. However, the specific contributions proposed by each module we proposed are still unknown. In this subsection, we conduct an ablation study to audit the performance gain provided by the two proposed modules. Concretely, we respectively remove the probabilistic variation assessment (PVA) and probabilistic variation calibration (PVC) that we introduced in Sec. 3.2 and Sec. 3.3. The results are represented in Table 3, where each module contributes a certain improvement to our proposed method. Besides, the PVC approach seems to play a more critical role, which can still serve as a valid adversary without the PVA. Thus, in practical scenarios, we can consider removing the PVA to reduce the frequency of accessing public APIs.

**Table 4: The performance of SPV-MIA against LLMs fine-tuned with DP-SGD w.r.t different noise magnitudes $\sigma_l$.**

| Privacy Budget $\epsilon$ | 1 | 2 | 4 | +inf |
|---|---|---|---|---|
| Wiki | 0.785 | 0.832 | 0.875 | 0.951 |
| AG News | 0.766 | 0.814 | 0.852 | 0.903 |
| Xsum | 0.771 | 0.827 | 0.867 | 0.937 |
| Avg. | 0.774 | 0.824 | 0.865 | 0.930 |

## 4.6 Defending against MIAs

As privacy risks emerge from various attacks, including data extraction attack [7], model extraction attack [18], and membership inference attack [36, 55, 67], the research community actively promotes defending methods against these attacks [24, 40]. DP-SGD [1] is one of the most widely adopted defense methods based on differential privacy [13] to provide mathematical privacy guarantees. Through DP-SGD, the amount of information the parameters have about a single data record is bound. Therefore, the privacy leaked from the target model will not exceed the upper bound, regardless of how many outputs we obtain from the target model. Specifically, DP-SGD is realized by adding noise to the clipped gradients:

$$\tilde{\boldsymbol{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \text{clip}\left(\boldsymbol{g}_t\left(x_i\right), C\right) + \mathcal{N}\left(0, \sigma^2 C^2 \boldsymbol{I}\right)\right), \tag{12}$$

where $C$ is the clipping norm, $\sigma$ denotes the noise scale, and $L$ represents the group size. We follow the same manner as the existing study [30] and train LLaMA with DP-Adam on the aforementioned three datasets. The results are summarized in Table. 4, where demonstrate DP-SGD can reduce the privacy risk with a certain. However, an excessively high privacy budget can lead to a performance degradation of the LLM. Under a moderate privacy budget, SPV-MIA still poses a significant risk of privacy leakage.

# 5 RELATED WORKS

## 5.1 Large Language Models

In the past year, large language models (LLMs) have dramatically improved performances on multiple natural language processing (NLP) tasks and consistently attracted attention in both academic and industrial circles [38]. The existing LLMs primarily fall into three categories: causal language modeling (CLM) (e.g. GPT), masked language modeling (MLM) (e.g. BERT), and Sequence-to-Sequence (Seq2Seq) approach (e.g. BART). Among these LLMs, CLMs have achieved the dominant position with the exponential improvement of model scaling [72]. Therefore, we select CLM as the representative LLM in this work for evaluation. As LLMs can absorb extensive

knowledge from pre-trained on large-scale corpora, they have the potential to serve as area experts via fine-tuning on specific domains [11]. Thus, LLMs are of great value for massive applications across multiple domains, such as finance [65], education [9], healthcare [66] and scientific research [57]. The widespread usage of LLMs has led to much other contemporaneous work on quantifying the privacy risks of LLMs [36, 42, 48]. Some preliminary works attempt to capture sensitive information, such as telephone numbers and postcodes, via elaborately designed prompting [29, 48]. In this work, we audit privacy leakages of LLMs through distinguishing whether or not a specific data record is used for fine-tuning the target LLM.

## 5.2 Membership Inference Attack

Membership inference attack (MIA) was firstly introduced in machine learning models by Shokri et al. [55], which aim to estimate the probability of a specific data sample was utilized in the training set of a machine learning model. Initially, following Shokri et al., most of the work was focused on the most common classification tasks in machine learning [6, 8, 34]. With the rapid development of other machine learning tasks, such as recommendation and generation tasks, MIAs against these task-specific models became a research direction of great value, and have been well investigated [12, 16, 70]. Meanwhile, the chatbot ChatGPT released by OpenAI has propelled the attention towards LLMs to the peak over the past year, which promotes the study on MIA against language models (LMs). Mireshghallah et al. [39] proposed the first MIA, Likelihood Ratio Attack (LiRA), against MLMs via adopting pre-trained models as reference model. Following this study, Mireshghallah et al. [41] further adapted LiRA for CLMs. Furthermore, Mattern et al. [36] pointed out the unrealistic assumption of a reference model trained on similar data, then substitute it with a neighbourhood comparison method. However, there is still a gap for MIAs against LLMs. Although MIAs against LMs have been studied by several works, the attack performance of existing MIAs in regard to LLMs that with large-scale parameters and pre-trained on tremendous corpora is still not clear. Therefore, we evaluate previous MIAs on LLMs in practical scenarios, and reveal that they are impracticable on LLMs due to their strict requirements and over-optimistic assumptions. Then, we propose a Membership Inference Attack based on Self-calibrated Probabilistic Variation (SPV-MIA), which disclose significant privacy risks on practical LLM applications.

# 6 CONCLUSION

In this paper, we reveal the unsatisfying performances of existing MIA methods against LLMs for practical applications and interpret this phenomenon from two perspectives. First, existing MIAs heavily rely on overfitting in the target LLM, which is usually avoided before releasing LLM for public access. Second, reference-based attacks seem to pose impressive privacy leakages by comparing the sampling probabilities of the target record between target and reference LLMs, but the inaccessibility of the appropriate reference dataset will be a big obstacle to deploying it in practice. To address these limitations, we propose a Membership Inference Attack based on Self-calibrated Probabilistic Variation (SPV-MIA), where we introduce a more reliable membership signal based on memorization rather than overfitting, then we propose a self-prompt approach to extract reference dataset from LLM itself in a practical manner. We conduct substantial experiments to validate SPV-MIA with state-of-the-art baselines across multiple representative LLMs. The results represent the superiority of SPV-MIA over all baselines and verify its effectiveness in extreme conditions.

# REFERENCES


[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*. Association for Computing Machinery, New York, NY, USA, 308–318.

[2] John Abascal, Stanley Wu, Alina Oprea, and Jonathan Ullman. 2023. TMI! Finetuned Models Leak Private Information from Their Pretraining Data. arXiv:2306.01181 [cs]

[3] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. 2023. *Falcon-40B: an open large language model with state-of-the-art performance*. Technical Report. Technical report, Technology Innovation Institute.

[4] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. arXiv:2310.05130 [cs]

[5] Andrew P Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, 7 (1997), 1145–1159.

[6] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. 2022. Membership Inference Attacks From First Principles. In *2022 IEEE Symposium on Security and Privacy (SP)*. 1897–1914.

[7] 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)*. 2633–2650.

[8] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-Only Membership Inference Attacks. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 1964–1974.

[9] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*. 323–325.

[10] Gerrit J. J. Van den Burg and Chris Williams. 2021. On Memorization in Probabilistic Deep Generative Models. In *Advances in Neural Information Processing Systems*.

[11] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. *Nature Machine Intelligence* 5, 3 (March 2023), 220–235.

[12] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. 2023. Are Diffusion Models Vulnerable to Membership Inference Attacks?. In *Proceedings of the 38th International Conference on Machine Learning, {ICML} 2023*. PMLR. arXiv:2302.01316 [cs]

[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.

[14] Vitaly Feldman. 2020. Does Learning Require Memorization? A Short Tale about a Long Tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC 2020)*. Association for Computing Machinery, New York, NY, USA, 954–959.

[15] Vitaly Feldman and Chiyuan Zhang. 2020. What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation. *Advances in Neural Information Processing Systems* 33 (2020), 2881–2891.

[16] Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2023. A Probabilistic Fluctuation based Membership Inference Attack for Diffusion Models. *arXiv e-prints* (2023), arXiv–2308.

[17] Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. 2023. Large language model AI chatbots require approval as medical devices. *Nature Medicine* (2023), 1–3.

[18] Xuanli He, Lingjuan Lyu, Lichao Sun, and Qiongkai Xu. 2021. Model Extraction and Adversarial Transferability, Your BERT is Vulnerable!. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2006–2012.

[19] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems* 28 (2015).

[20] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[21] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. *Comput. Surveys* 54, 11s (Sept. 2022), 235:1–235:37.

[22] Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, Xing Xu, and Soujanya Poria. 2023. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. arXiv:2304.01933 [cs]

[23] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[24] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. 2019. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 259–274.

[25] Belveze Jules. 2022. TLDR News Dataset. https://huggingface.co/datasets/JulesBelveze/tldr_news

[26] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[27] Mario Michael Krell, Matej Kosec, Sergio P Perez, and Andrew Fitzgibbon. 2021. Efficient Sequence Packing without Cross-contamination: Accelerating Large Language Models without Impacting Performance. *arXiv preprint arXiv:2107.02027* (2021).

[28] Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight Poisoning Attacks on Pretrained Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2793–2806.

[29] Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. 2021. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? arXiv:2104.07762 [cs]

[30] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2021. Large Language Models Can Be Strong Differentially Private Learners. In *International Conference on Learning Representations*.

[31] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4582–4597.

[32] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems* 35 (2022), 1950–1965.

[33] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. GPT understands, too. *AI Open* (2023).

[34] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. 2022. Membership Inference Attacks by Exploiting Loss Trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*. Association for Computing Machinery, New York, NY, USA, 2085–2098.

[35] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

[36] 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:2305.18462 [cs]

[37] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*.

[38] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *Comput. Surveys* 56, 2 (Sept. 2023), 30:1–30:40.

[39] Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 8332–8347.

[40] Fatemehsadat Mireshghallah, Huseyin Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. 2021. Privacy Regularization: Joint Privacy-Utility Optimization in LanguageModels. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3799–3807.

[41] Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022. An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1816–1826.

[42] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. In *Proceedings of the 38th International Conference on Machine Learning, {ICML} 2023*.

[43] Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1797–1807.

[44] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *2019 IEEE Symposium on Security and Privacy (SP)*. 739–753.

[45] OpenAI. 2023. ChatGPT: Optimizing Language Models for Dialogue. http://web.archive.org/web/20230109000707/https://openai.com/blog/chatgpt/.

[46] OpenAI. 2023. OpenAI Documentation-Text Generation-Completions API . https://platform.openai.com/docs/guides/text-generation/completions-api.

[47] Yin Minn Pa Pa, Shunsuke Tanizaki, Tetsui Kou, Michel Van Eeten, Katsunari Yoshioka, and Tsutomu Matsumoto. 2023. An Attacker's Dream? Exploring the Capabilities of ChatGPT for Developing Malware. In *Proceedings of the 16th Cyber Security Experimentation and Test Workshop*. 10–18.

[48] Charith Peris, Christophe Dupuy, Jimit Majmudar, Rahil Parikh, Sami Smaili, Richard Zemel, and Rahul Gupta. 2023. Privacy in the Time of Language Models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 1291–1292.

[49] Lutz Prechelt. 2002. Early stopping-but when? In *Neural Networks: Tricks of the trade*. Springer, 55–69.

[50] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[51] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[52] Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, and German Rigau. 2010. Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/222_Paper.pdf

[53] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Herve Jegou. 2019. White-Box vs Black-box: Bayes Optimal Strategies for Membership Inference. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 5558–5567.

[54] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed Systems Security (NDSS) Symposium 2019*.

[55] 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)*. 3–18.

[56] James Stewart. 2001. *Multivariable calculus: concepts and contexts*. Brooks/Cole.

[57] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science. arXiv:2211.09085 [cs, stat]

[58] Yingjie Tian and Yuqi Zhang. 2022. A comprehensive survey on regularization strategies in machine learning. *Information Fusion* 80 (2022), 146–166.

[59] Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 38274–38290.

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

[61] Priyan Vaithilingam, Tianyi Zhang, and Elena L Glassman. 2022. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Chi conference on human factors in computing systems extended abstracts*. 1–7.

[62] Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed Data Poisoning Attacks on NLP Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 139–150.

[63] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

[64] Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. 2022. On the Importance of Difficulty Calibration in Membership Inference Attacks. In *International Conference on Learning Representations*.

[65] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564 [cs, q-fin]

[66] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A Large Language Model for Electronic Health Records. *npj Digital Medicine* 5, 1 (Dec. 2022), 1–9.

[67] 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 (CCS '22)*. Association for Computing Machinery, New York, NY, USA, 3093–3106.

[68] 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)*. IEEE, 268–282.

[69] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially Private Fine-tuning of Language Models. In *International Conference on Learning Representations*.

[70] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. 2021. Membership Inference Attacks Against Recommender Systems. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)*. Association for Computing Machinery, New York, NY, USA, 864–879.

[71] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems* 28 (2015).

[72] 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. 2023. A Survey of Large Language Models. arXiv:2303.18223 [cs]

# A APPENDIX

## A.1 Notations of This Work

**Table 5: Notations and descriptions.**

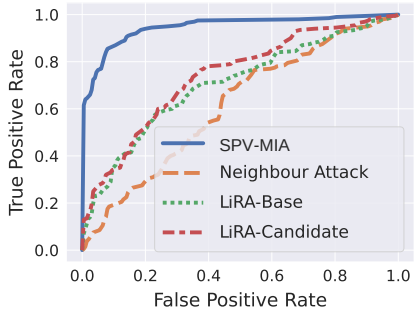
| Notation | Description |
|---|---|
| $\boldsymbol{x}$ | A specific data record. |
| $\widetilde{\boldsymbol{x}}_n$ | A paraphrasing text record of the target text record $\boldsymbol{x}$. |
| $D_{mem}$ | The training dataset utilized for LLM fine-tuning. |
| $D_{non}$ | A disjoint dataset from the training dataset. |
| $D_{refer}$ | The reference dataset that collected for fine-tuning reference LLM. |
| $m^{(j)}$ | The membership of the data record $\boldsymbol{x}^{(j)}$, 1 represents member, whereas 0 represents non-member. |
| $\theta$ | The parameters of the target large language model (LLM). |
| $\phi$ | The parameters of the reference LLM. |
| $\psi$ | The parameters of the paraphrasing LLM. |
| $\mathcal{A}(\boldsymbol{x}, \theta)$ | The adversary algorithm for MIA. |
| $p_\theta(\boldsymbol{x})$ | The probability of text record $\boldsymbol{x}$ being sampled by the generative model $\theta$ . |
| $p_\theta(\widetilde{\boldsymbol{x}}_n)$ | The probability of paraphrasing text $\widetilde{\boldsymbol{x}}_n$ being sampled by the generative model $\theta$ . |
| $\Delta p_\theta(\boldsymbol{x})$ | The calibrated probability of text record $\boldsymbol{x}$. |
| $\widetilde{p}_\theta(\boldsymbol{x})$ | The probabilistic variation of $\boldsymbol{x}$ measured on the target LLM $\theta$. |
| $\widetilde{p}_\phi(\boldsymbol{x})$ | The probabilistic variation of $\boldsymbol{x}$ measured on the reference LLM $\phi$. |
| $\Delta\widetilde{p}_{\theta,\phi}(\boldsymbol{x})$ | The calibrated probabilistic variation of $\boldsymbol{x}$ measured on both the target LLM $\theta$ and the reference LLM $\phi$. |
| $q_\psi(\cdot, \boldsymbol{x})$ | The paraphrasing function parameterized by the paraphrasing LLM $\psi$. |
| $N$ | The query times for estimating $\widetilde{p}_\theta(\boldsymbol{x})$. |

## A.2 Detailed Information for Reproduction

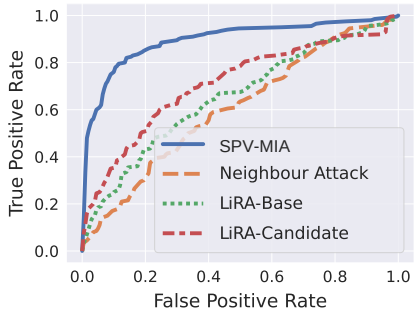
For each dataset, we pack multiple tokenized sequences into a single input, which can effectively reduce computational consumption without sacrificing performance [27]. Besides, the packing length is set to 128 tokens. Then, we use 10,000 samples for fine-tuning over pre-trained LLMs and 1,000 samples for evaluation. The detailed information of datasets is summarized in Tab. 6. For each target LLM, we let it fine-tuned with the training batch size of 16, and trained for 10 epochs. The learning rate is set to 0.0001. We adopt the AdamW optimizer [35] to achieve the generalization of LLMs, which is composed of the Adam optimizer [26] and the L2 regularization. For GPT-2, which has a relatively small scale, we adopt the full fine-tuning, which means all parameters are trainable. For other LLMs that are larger, we utilize a parameter-efficient fine-tuning method, Low-Rank Adaptation (LoRA) [20], as the default fine-tuning method. For paraphrasing text, we follow the setting in DetectGPT [4] randomly mask 15% tokens within a text and generate 20 paraphrased texts for each target text record. For the reference LLM fine-tuned with our proposed self-prompt approach, we utilize the domain-specific data as the default prompt text source. Then, we collect 10,000 generated texts from target LLMs with an equal length of 128 tokens to construct reference datasets. We fine-tune the reference LLM for 4 epochs and the training batch size of 16.

## A.3 Supplementary Experiments Results

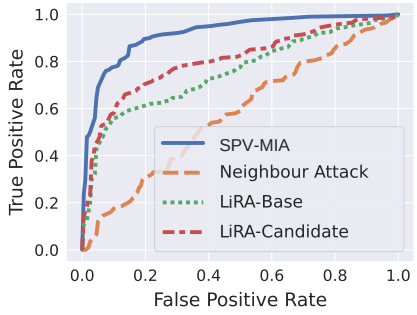
As a supplement to the main experimental results represented in Tab. 1, we further provide the raw ROC curve for a more comprehensive presentation in Fig. 5.




(a) Wiki




(b) AG News




(c) Xsum

**Figure 5: ROC curves of SPV-MIA and the top-three best baselines on LLaMAs fine-tuned over three datasets.**

**Table 6: Detailed split and other information of datasets.**

| Dataset | Relative Datasets | | Target Model | | Reference Model | |
|---|---|---|---|---|---|---|
| | Domain-specific | Irrelevant | # Member | # Non-member | # Member | # Non-member |
| Wikitext-103 | Wikicorpus | AG News | 10,000 | 1,000 | 10,000 | 1,000 |
| AG News | TLDR News | Xsum | 10,000 | 1,000 | 10,000 | 1,000 |
| Xsum | CNNDM | Wikitext-103 | 10,000 | 1,000 | 10,000 | 1,000 |