DF-MIA: A Distribution-Free Membership Inference Attack on Fine-Tuned Large Language Models

Zhiheng Huang[1,2*], **Yannan Liu**[2], **Daojing He**[1], **Yu Li**[1,3†]
[1]Harbin Institute of Technology, Shenzhen, China
[2]ByteDance, China,
[3]Zhejiang Univerisity, China
huangzhiheng@stu.hit.edu.cn, liuyannan.space@bytedance.com
hedaojinghit@163.com, yu.li.sallylee@gmail.com

Abstract

Membership Inference Attack (MIA) aims to determine if a specific sample is present in the training dataset of a target machine learning model. Previous MIAs against fine-tuned Large Language Models (LLMs) either fail to address the unique challenges in the fine-tuned setting or rely on strong assumption of the training data distribution. This paper proposes a distribution-free MIA framework tailored for fine-tuned LLMs, named DF-MIA. We recognize that samples await to test can serve as a valuable reference dataset for fine-tuning reference models. By enhancing the signals of non-member samples within this reference dataset, we can achieve a more reliable and practical calibration of probabilities, improving the differentiation between members and non-members. Leveraging these insights, we have developed a two-stage framework that employs specially designed data augmentation and perturbation techniques to prioritize the significance of non-members and mitigate the influence of potential members within the reference dataset. We evaluate our method on three representative LLM models ranging from 1B to 8B on three datasets. The results demonstrate that the DF-MIA significantly enhances the performance of MIA.

1 Introduction

Large language models (LLMs) have demonstrated exceptional capabilities across a variety of tasks, including code generation (Vaithilingam, Zhang, and Glassman 2022), machine translation (Zhang, Haddow, and Birch 2023), and news summarization (Xiao and Chen 2023). Due to the extensive datasets and significant computational resources required for their training, the pretraining-finetuning paradigm has become the leading methodology for crafting domain-specific models (Thangarasa et al. 2023). This approach is further facilitated by the accessibility of open-source pretrained models like GPT-2 (Radford et al. 2019) and LLaMA (Touvron et al. 2023). Notably, products such as OpenAI's ChatGPT (Yetiştiren et al. 2023) and GitHub Copilot (Chen et al. 2021) are developed within this framework, highlighting its effectiveness and impact.

However, existing studies have revealed that, despite the powerful performance, LLMs possess substantial memory capabilities and are susceptible to potential privacy risks (Peris et al. 2023; Ishihara 2023; Carlini et al. 2023; Zeng et al. 2024). A notable technique that exploits this memory capacity for privacy leakage is the Membership Inference Attack (MIA) (Mireshghallah et al. 2022b; Shokri et al. 2017). This attack specifically aims to determine whether a particular record belongs to the training set of the target model, thereby compromising privacy (Mattern et al. 2023).

Existing MIAs can be broadly divided into two categories: *reference-free* and *reference-based* strategies. *Reference-free* methods primarily rely on the target model's output predictions for membership inference (Song and Mittal 2021; Yeom et al. 2018; Mattern et al. 2023). For example, the Loss Attack method uses the loss derived from the output, while Min-%k leverages a subset of output probabilities. However, these methods often struggle with varying sample difficulties. Specifically, low losses and high probabilities nonmembers tend to be misclassified as members, resulting in false positives (FP). Conversely, high losses and low probabilities members are often misclassified as nonmembers, leading to false negatives (FN). To address these challenges, *reference-based* methods have been developed (Mireshghallah et al. 2022b; Mattern et al. 2023). These approaches use a reference model to assess sample difficulty and adjust the original judgment accordingly. For precise calibration, the reference model is typically trained on datasets with a similar distribution to the target model's training data.

However, existing MIAs encounter significant challenges in fine-tuned LLM scenarios where privacy concerns intensify due to the use of domain-specific confidential data (Mireshghallah et al. 2022b; Zeng et al. 2024). These attacks either fail to meet the unique requirements of fine-tuned LLMs or rely on assumptions about training data distribution. For example, non-members misclassified as members may belong to the members of the pretraining dataset, thus leading to low-loss and erroneous classifications. Furthermore, acquiring the training data distribution needed for reference-based methods proves to be impractical (Mattern et al. 2023; Carlini et al. 2022).

To address the challenges above, this work introduces a distribution-free MIA framework, named *DF-MIA*, specif-

[*] Contribution during internship at ByteDance.

[†] Corresponding author.

ically designed for fine-tuned LLMs. Unlike prior approaches, we do not rely on assumptions about the training data distribution to craft a high-quality reference model. Our strategy hinges on fine-tuning the reference model using *samples to test*, i.e., samples pending classification as members or non-members. This approach enables the reference model to generate low-loss values for non-members, facilitating more accurate calibration. However, the potential members in samples to test can complicate the calibration. To mitigate this, we leverage techniques from previous attacks (such as the Loss Attack) to estimate the likelihood of a sample being a non-member and emphasize their importance. Additionally, we recommend using the base model—on which the target model is fine-tuned—as the initial reference model. This allows us to better identify whether a sample comes from the fine-tuning dataset, reducing the distraction caused by the pre-training dataset.

Overall, our contributions can be summarized as follows:

- We propose a novel distribution-free MIA framework for fine-tuned LLMs. Unlike the previous method, we chose the base model as the initial reference model and fine-tune it without any assumption on the training data distribution.
- To fine-tune the reference model, we utilize samples to test that contain nonmember samples to construct a reference dataset. This dataset is readily available and requires no extra effort.
- To mitigate the influence of members in the reference dataset, we introduce a two-stage framework. We 1) utilize data augmentation on the dataset to assign higher weights for nonmember samples and 2) apply data perturbation on high-weight samples to weaken the influence of high-loss member samples.

We conduct extensive experiments to validate the performance of DF-MIA. Our result indicates that the DF-MIA outperforms existing MIAs and reveals a significant privacy threat to fine-tuned LLMs. Our code is available at https://github.com/HZHKevin/DF-MIA.

2 Related Work

Large Language Model and its Memorization The Large Language Model (LLM) has attracted widespread attention in both industrial and academic circles since its emergence (Wang 2024; Minaee et al. 2024). Leveraging extensive knowledge gained from pretraining on large-scale corpora, the LLM demonstrates its potential to become domain experts through fine-tuning on domain-specific datasets. (Ding et al. 2023). However, prior researches have identified that the LLM possesses a substantial memory capacity (Carlini et al. 2023, 2021; Yu et al. 2023; Nasr et al. 2023). Current studies suggest that this ability can be utilized to leak sensitive information from LLMs, leading to severe privacy risk (Aditya et al. 2024; Zeng et al. 2024; Mireshghallah et al. 2022b; Carlini et al. 2021).

Membership Inference Attack on LLM Membership Inference Attack (MIA) is exactly one of the threatening attacks that can utilize the memory capability of LLM to compromise privacy (Mireshghallah et al. 2022b). The main idea behind Membership Inference Attacks (MIAs) is to determine if a specific record exists in the training set of the target model (Shokri et al. 2017). This attack can compromise user privacy and has become a prevalent technique for evaluating privacy risks in the field of machine learning (Yeom et al. 2018; Mattern et al. 2023). MIAs against LLMs can be categorized into two branches: the black-box attack and the white-box attack. The white-box attack assumes partial or complete access to the target model (Li et al. 2023; Liu et al. 2022; Wang et al. 2024; Sablayrolles et al. 2019).

In contrast, for a black-box scenario, the adversary is solely permitted to obtain the responses of models, such as log probabilities (Sablayrolles et al. 2019). This condition is more realistic and consistent with practical application circumstances. The existing MIAs developed for LLMs can be categorized into two types: reference-free and reference-based attacks (Mattern et al. 2023).

The reference-free attacks only utilize the information of the target model. Loss Attack directly utilizes model loss information and classifies high loss samples as nonmembers (Mattern et al. 2023). Another approach, known as Min-%k, utilizes a simpler method that focuses solely on the negative log-probability of the bottom k% tokens with the lowest probabilities (Shi et al. 2024). Neighbor attack suggests that nonmembers are more likely to exhibit larger loss discrepancies with their text neighbor generated by perturbation (Mattern et al. 2023).

The reference-based attacks demonstrate strong performance but require the attacker to have access to the training data distribution of the target model. The Likelihood Ratio Attack (LiRA) is a method that introduces calibrated probabilities for samples to test by accessing a reference model, which is trained from a pretrained model with data that shares the similar distribution to the training set (Mireshghallah et al. 2022a).

Compared to the pretraining phase, fine-tuning often involves the use of domain-specific and private data, the leakage of which can seriously violate user privacy or intellectual property rights (Zeng et al. 2024). However, there are two main challenges for existing frameworks utilized in fine-tuned LLM scenarios: first, accessing the overall distribution of training data to acquire high-quality reference models can be challenging in the realistic scenario, since it is impossible to accurately represent the entire distribution without having complete access to the training dataset (Mattern et al. 2023; Carlini et al. 2022). Second, partial low-loss nonmembers may come from the pretraining dataset rather than the fine-tuning dataset. Our attack do not rely on data distribution information, and directly utilize samples to test to generate the loss calibration for nonmember samples.

3 Methodology

In this section, we introduce an innovative membership inference attack against fine-tuned LLMs.

3.1 Framework

The fundamental goal of MIAs is to capture indicators of overfitting and discern whether a sample is a member of the

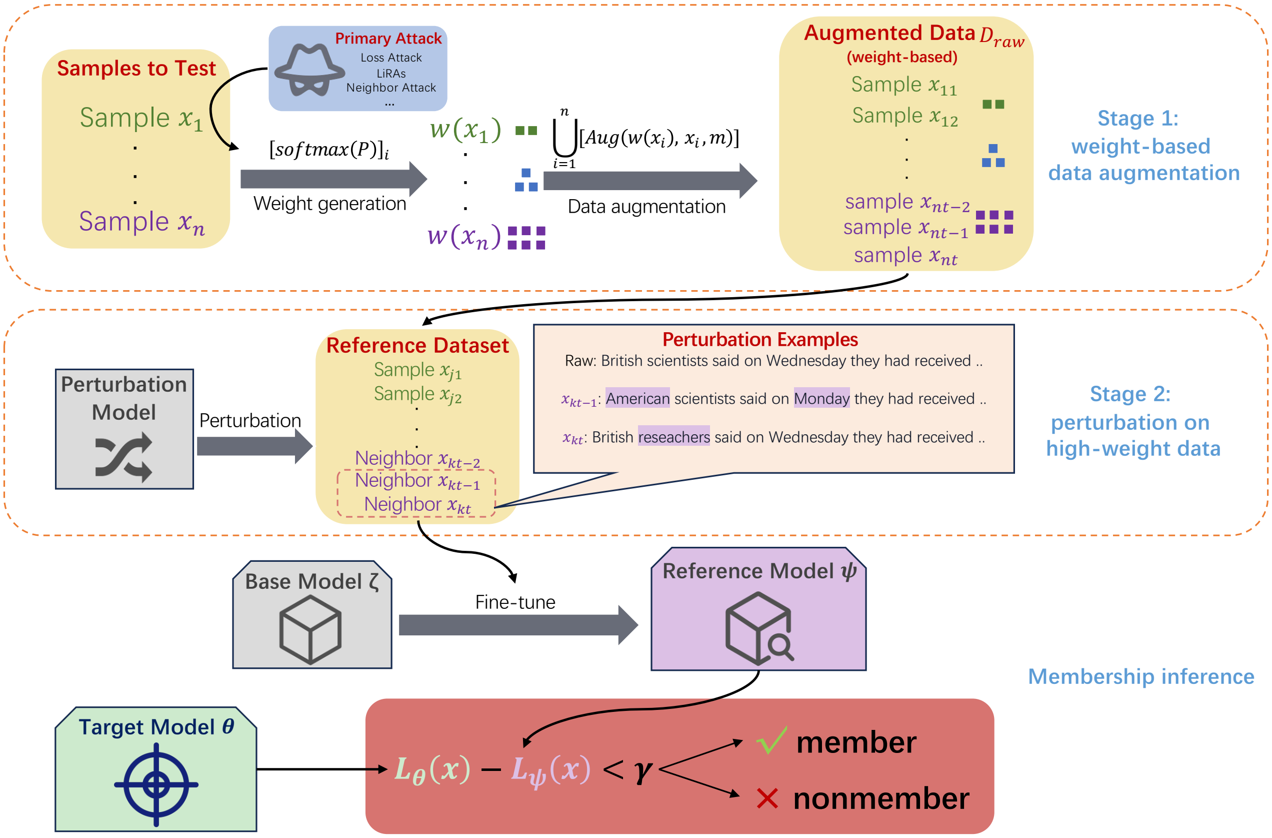


Figure 1: The overview of our framework. We utilize a two-stage approach to generate the reference dataset. Then, we fine-tune the base model with the reference dataset and eventually deploy the reference-based membership inference attack.

training set (Yeom et al. 2018). The most straightforward MIA framework can be summarized as:

$$A_\theta(x) = \mathbb{I}[L(\theta, x) < \gamma], \tag{1}$$

where $L(\theta, x)$ denotes the model loss for a specific sample x, while γ represents a predetermined threshold. The indicator function $\mathbb{I}$ classifies samples as members if their $L(\theta, x)$ values are lower than γ. However, the effectiveness of this method is hindered by the sample difficulty. On the one hand, for easy inputs, the model tends to give a low loss even if these inputs are non-members, resulting in false positives (FP). On the other hand, for difficult inputs, the model is likely to assign a higher loss even if these inputs are indeed members, resulting in false negatives (FN). Therefore, many studies propose calibrating the loss value with a difficulty score $d(x)$ (Mattern et al. 2023; Mireshghallah et al. 2022b). The core idea of these studies can be formulated as:

$$A_\theta(x) = \mathbb{I}[L(\theta, x) - d(x) < \gamma]. \tag{2}$$

Despite their effectiveness, the current calibration methods still face challenges in the context of fine-tuned LLMs. First, existing approaches do not explicitly distinguish between the pre-training and fine-tuning datasets. This can lead to falsely identified members that belong to the members of the often open-source pertaining dataset instead of the private fine-tuning dataset. Second, to achieve high efficacy, existing reference-based frameworks often assume access to a reference dataset whose data distribution is similar to the training set. However, obtaining accurate data distribution without complete access to the training set can be challenging in real-world scenarios (Mattern et al. 2023).

Our work addresses the above challenges by considering the dataset difference and eliminating the reliance on the training data distribution assumption. Our framework stems from a simple intuition: for a given $L(\theta, x)$, the FP and FN samples can be better distinguished with a higher $d(x)$ value for members or a lower $d(x)$ value for nonmembers. We observe that directly fine-tuning the base model with nonmember samples can effectively lower $d(x)$ value for them in the reference model. This motivate us to utilize samples to test to construct a reference dataset since it contains all nonmember samples. However, member samples also exist in samples to test which are noises. Fortunately, we find that the outcome $L(\theta, x)$ or $L(\theta, x) - d(x)$ from a primary attack (e.g. Loss Attack, Neighbor Attack) can serve as an indicator of the confidence level associated with nonmembers. Then, the weights for each samples are generated based on their indicator values, and higher weights are assigned to samples which are more likely to be nonmembers. As high weights may also be assigned to FN samples, we apply perturbation to reduce their influence. We eventually fine-tune the base model with augmented and perturbated samples to test and gain the reference model. The difficulty score of each sample is represented by its loss value of the reference model. With this difficulty score to calibrate, we can take all samples into account without neglecting the FP which come from the pretraining dataset. Besides, we do not rely on any data distribution information in this framework.

As illustrated in Figure 1, our reference dataset is generated by two stages. In the first stage, the indicator value for each sample is generated from a primary attack. Then, the weights for each samples are generated based on their indicator values. We employ weight-based data augmentation with replication on samples to test to enhance the weight of nonmembers. In the second stage, we employ perturbation on the acquired data which gain relative high weight. Subsequently, having obtained the reference dataset, we finetune it with the base model and proceed to implement the reference-based membership inference.

3.2 Stage 1: Weight-based Data Augmentation

Our target is to enhance weights of nonmembers in the reference dataset. We find that the indicator value $p(x) = L(\theta, x)$ or $p(x) = L(\theta, x) - d(x)$ from a primary attack can serve as signals of the confidence level associated with nonmembers. After obtaining indicator values $P = \{p(x_1), \ldots, p(x_n)\}$ from samples to test, we then use a $softmax$ function to normalize and calculate corresponding weight $w(x_i) = [softmax(P)]_i$ for each sample. With the obtained weights, we can utilize data augmentation to generate the raw reference dataset from samples to test as follows:

$$D_{raw} = \bigcup_{i=1}^{n} [Aug(w(x_i), x_i, m)] \tag{3}$$

where m represents the preset volume of augmented data. Each sample will be copied $m \times w(x_i)$ times to generate the augmented data D_{raw}.

3.3 Stage 2: Perturbation on High-weight Data

Based on our intuition, we expect nonmember samples to obtain higher weights. However, partial member samples can also obtain high weights due to FN. We observed that the fundamental difference between nonmembers and FN lies in whether they have been fine-tuned on the base model. This observation supports the application of perturbation to amplify their differences. Specifically, nonmembers with high weights can be divided into two parts: a section that is easy to fine-tune, another segment that akin to FN and is hard to fine-tune. The fine-tuning of sample neighbors has a limited impact on easily fine-tuned nonmembers, while posing an obstacle to reducing $d(x)$ for hard fine-tuning nonmembers and FN. This aids in better distinguishing easily fine-tuned non-members. Accordingly, we apply perturbation to high-weight samples within the augmented data in order to generate the reference dataset.

3.4 Membership Inference with Reference Model

The framework of our attack can be summarized in Algorithm 1. For each sample x_i, we first generate its indicator value $p(x_i)$ with the primary attack. With the obtained indicators P, we use $softmax$ to generate weight $w(x_i)$ for each sample x_i. The replication times in augmented dataset D_{raw} for each sample x_i is calculated by $m \times w(x_i)$. Then, we obtain the reference dataset N by applying perturbation on top-r% high-weight data. After acquiring the reference dataset, we fine-tune it on the base model for the purpose of conducting the membership inference attack with obtained reference model. In contrast to existing reference-based frameworks, the DF-MIA generates $d(x)$ from samples to test.

Algorithm 1: DF-MIA

```
Input: Samples $\{x_1, x_2, ..., x_n\}$, target model $\theta$, base
    model $\zeta$, primary attack method $Attk()$, data augmentation
    algorithm $Aug$, volume of the reference dataset
    $m$, perturbation model $F$, perturbation rate $r$, fine-tune
    method $\Omega$, threshold $\gamma$
Output: Results of samples $\{res_1, res_2, ..., res_n\}$, where
    $res_i = 1$ represent $x_i$ is a member, otherwise $x_i$ is a
    nonmember
$P \leftarrow \{\}, , D_{raw} \leftarrow \{\}, N \leftarrow \{\}$
for each $x_i \in \{x_1, x_2, ..., x_n\}$ do
    $p(x_i) \leftarrow Attk(x_i)$
    $P \leftarrow P \cup p(x_i)$
end for
for each $p(x_i) \in P$ do
    $w(x_i) \leftarrow [softmax(P)]_i$
    $D_{raw} \leftarrow D_{raw} \cup Aug[(w(x_i), x_i, m)]$
end for
for each $n_i \in AD$ do
    $N \leftarrow N \cup F(n_i, r)$
end for
$\Psi \leftarrow \Omega(\zeta, N)$
for each $x_i \in \{x_1, x_2, ..., x_n\}$ do
    $res_i = \mathbb{I}[L(\theta, x_i) - L(\Psi, x_i) < \gamma]$
end for
return $\{res_1, res_2, ..., res_n\}$
```

4 Experiments

4.1 Experiment Setup

Datasets We evaluate our framework on three datasets from various domains: Wikitext-103, AGNews, and XSum. To be specific, the Wikitext-103 (Merity et al. 2017) contains academic writing summaries, the AGNews (Zhang, Zhao, and LeCun 2015) involves summaries of news topics, and the XSum (Narayan, Cohen, and Lapata 2018) contains document summaries.

Target Models To achieve a comprehensive evaluation, we conduct experiments on three widely used pre-trained LLMs, with parameter scales ranging from 1.3 billion to 8 billion:

- **GPT-Neo-1.3B** (Black et al. 2022): gpt-neo is an open-source replication of GPT-3 architecture.
- **OPT-2.7B** (Zhang et al. 2022): OPT is a LLM utilizing transformer architecture released by Meta AI.

- **LLaMA-3-8B** (AI@Meta 2024): In 2024, Meta AI open-sourced the highly esteemed LLM framework, LLaMA-3-8B.

To obtain the target models, we follow the method in (Mattern et al. 2023) and fine-tune the based models on each dataset described above. The detailed settings are described in the supplementary material.

Baselines We evaluate five representative black-box MIAs suitable for fine-tuned LLMs to comprehensively assess our proposed method, encompassing three reference-free attacks and two reference-based attacks.

- **Loss Attack** (Yeom et al. 2018): A commonly used MIA baseline involves identifying nonmembers by evaluating whether the loss values exceed a predetermined threshold.
- **Neighbour Attack** (Mattern et al. 2023): The Neighbour Attack employs the average loss of plausible neighbor texts as the difficulty score $d(x)$ to calibrate the loss value of the target model.
- **Min-k% Attack** (Shi et al. 2024): The method is founded on the premise that non-member samples are likely to contain more low-probability outlier tokens. Initially, this method was designed for application to pretrained models, but we have also utilized it for attack assessment in the context of fine-tuned scenarios.
- **LiRA-Base** (Mireshghallah et al. 2022b): A reference-based attack uses the loss values of the base model as difficulty scores to adjust the sample loss in order to execute membership inference.
- **LiRA-Candidate** (Ye et al. 2022): An alternative version of the LiRA-Base, which utilizes an openly accessible dataset from the similar distribution as the fine-tuning dataset to fine-tune the base model. The loss values of obtained reference model can provide a more reliable difficulty scores compared to LiRA-Base.

Evaluation Metrics Following previous works, we evaluate the MIA performance with two metrics:

- **AUC** (Shokri et al. 2017): The Area Under the ROC Curve (AUC) is a widely used metric that provides an overall assessment of MIA performance. The ROC curve itself is a plot that depicts the False Positive Rate (TPR) versus the True Positive Rate (FPR) across various classification thresholds.
- **TPR@1%FPR** (Carlini et al. 2022): The true positive rate at 1% false positive rate. Instead of paying equal attention to members and nonmembers, this metric pays more attention to members and evaluates whether the attacker can confidently identify members of the training dataset.

Hardware Platform Our experiments are conducted using 4 NVIDIA A800 GPUs.

4.2 Compare Our Method with Baselines

We report the AUC and TPR@1%FPR results of DF-MIA on three LLMs across three datasets. In terms of AUC values, as shown in Table 1, we find that DF-MIA outperforms all baselines. Compared to our primary attack (Neighbor Attack), DF-MIA achieves better performance by over 20%. It is better than LiRA-Candidate which utilizes distribution information in all experimental conditions. Regarding TPR@1%FPR in Table 2, we find that DF-MIA also outperforms all baselines. As a metric associated with risks at a relatively low threshold, a higher TPR@1%FPR also suggests a reduction of FP. We also exhibit Receiver Operating Characteristic (ROC) curves for DF-MIA and the five representative baselines in GPT-Neo-1.3B for the AGNews dataset to provide a more understandable visualization in Figure 2.

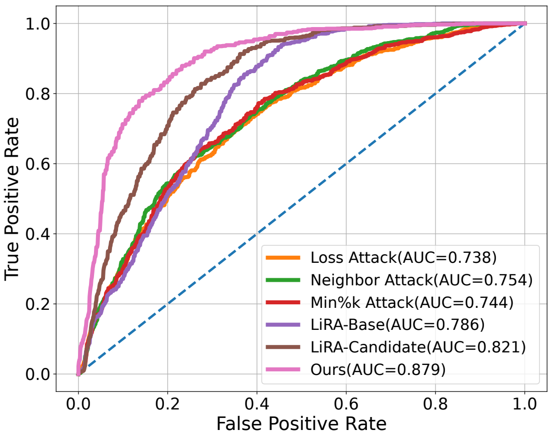


Figure 2: ROC curves of DF-MIA and the baselines. The target model is GPT-Neo-1.3B fine-tuned on AGnews dataset.

To further demonstrate that our framework capture nonmember signals and mitigate FP challenge, we extend the experiment on GPT-neo-1.3B using the AGnews dataset. As exhibited in Figure 3 (a) and (b), we demonstrate our hypothesis by observing framework performance with FP decreasing in the reference dataset. Specifically, we eliminate the augmentation on the top-k lowest confident nonmember. This variation continues to weaken the performance on both AUC and TPR@1%FPR, indicating that our framework can make better classification with nonmember signals. Compared to the decrement of FP, in Figure 3 (c) and (d) we notice that the decrement of FN further improve the AUC and TPR@1%FPR performance of DF-MIA . This suggests that FN poses a challenge to our framework and explains the rationale for applying perturbation to high-weight samples.

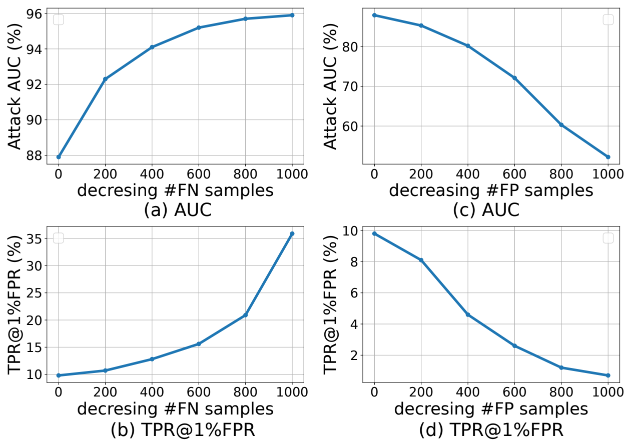
As a framework that allows for the selection of various primary attacks, we also present the enhanced performance with other attack options. The AUC and TPR@1%FPR results are shown in Figure 4. For AUC performance, a better primary attack may improve it, but the upper limit is still constrained by the FN. In terms of TPR@1%FPR enhancement, it is notably influenced by the TPR@1%FPR of its primary attack. It is argued that the ability to extract a tiny but high confidence subset of training data poses a severe privacy risk (Carlini et al. 2022), which emphasizes the threat of higher TPR@1%FPR. We have chosen the Neighbor Attack as our primary attack since its performance of is

Method	AGNews			Wiki			XSum		
	GPT-Neo	OPT	LLaMA	GPT-Neo	OPT	LLaMA	GPT-Neo	OPT	LLaMA
Loss Attack	0.738	0.611	0.611	0.671	0.637	0.585	0.730	0.720	0.615
Neighbor Attack	0.754	0.666	0.634	0.675	0.672	0.598	0.765	0.733	0.628
Min-k% Attack	0.744	0.731	0.617	0.694	0.656	0.599	0.725	0.708	0.611
LiRA-Base	0.786	0.782	0.746	0.707	0.668	0.658	0.850	0.708	0.809
LiRA-Candidate	<u>0.821</u>	<u>0.813</u>	<u>0.767</u>	<u>0.743</u>	<u>0.726</u>	<u>0.696</u>	<u>0.879</u>	<u>0.794</u>	<u>0.821</u>
Ours	**0.879**	**0.884**	**0.836**	**0.849**	**0.839**	**0.810**	**0.903**	**0.929**	**0.896**

Table 1: AUC performance on three LLMs across three datasets for DF-MIA and five baseline methods. Bold and underline represent the best and the second best result within each column.

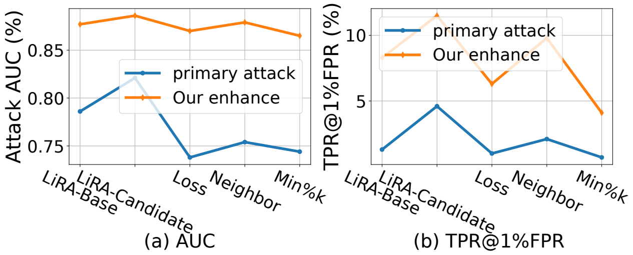
Method	AGNews			Wiki			XSum		
	GPT-Neo	OPT	LLaMA	GPT-Neo	OPT	LLaMA	GPT-Neo	OPT	LLaMA
Loss Attack	1.0%	1.4%	1.4%	1.5%	1.4%	1.6%	5.2%	5.5%	2.5%
Neighbor Attack	2.1%	2.1%	2.2%	1.9%	2.7%	2.0%	5.9%	8.3%	3.2%
Min-k% Attack	0.7%	2.5%	1.8%	1.8%	0.9%	1.9%	4.9%	6.3%	2.9%
LiRA-Base	1.3%	1.7%	2.6%	1.6%	2.0%	1.9%	9.8%	6.3%	8.1%
LiRA-Candidate	<u>4.6%</u>	<u>3.0%</u>	<u>3.2%</u>	<u>3.7%</u>	<u>3.3%</u>	<u>3.4%</u>	<u>11.3%</u>	<u>10.2%</u>	<u>9.1%</u>
Ours	**9.8%**	**10.2%**	**7.4%**	**5.9%**	**5.8%**	**8.4%**	**26.9%**	**22.8%**	**20.6%**

Table 2: TPR@1% FPR performance on three LLMs across three datasets for DF-MIA and five baseline methods. Bold and underline represent the best and the second best result within each column.



(a) AUC (c) AUC (b) TPR@1%FPR (d) TPR@1%FPR

Figure 3: Influence of eliminating weight-based data augmentation.



(a) AUC (b) TPR@1%FPR

Figure 4: The performance of DF-MIA when using different primary attacks.

relatively high without requiring distribution information.

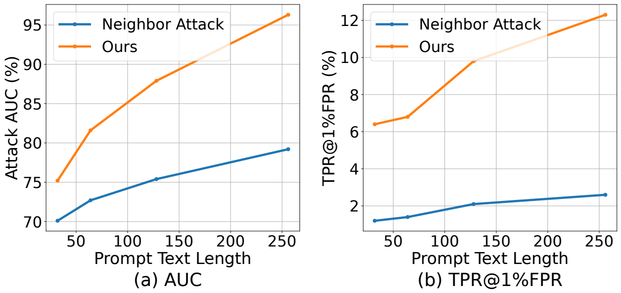
4.3 Impact of Hyperparameters

We analyze the impact of hyperparameters for our attack on GPT-neo-1.3B using the AGnews dataset.

Prompt Text Length: In Figure 5, we exhibit the results of our attack, comparing the primary attack (Neighbor Attack) across various text lengths (32, 64, 128, 256). The experimental result suggests that as the sample length increases, the effectiveness of attacks also increases.

Volume of the Reference Dataset: Figure 6 (a) and (b) illustrate the impact of the reference dataset volume, which is denoted as m. When the volume is 2000, it means that the data augmentation is not applied. We observe that the performance of our framework improves until $m > 10000$, after which it becomes inconsistent. Since there is a need for large enough data volume to amplify the weights of non-members in the reference dataset, highlighting the impact of our weight-based data augmentation. To strike a balance between efficiency and performance, we ultimately choose $m = 30000$.

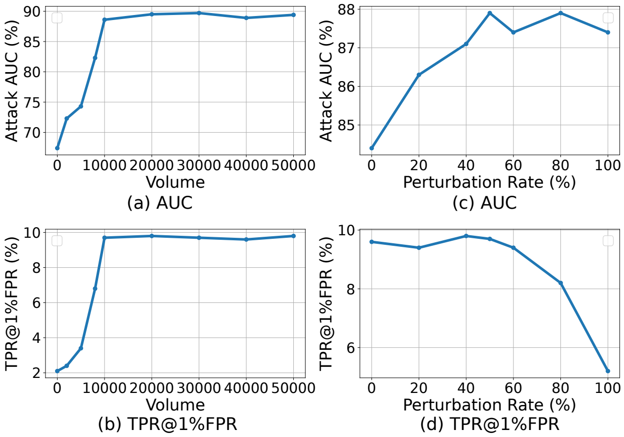
Perturbation Rate: Figure 6 (c) and (d) illustrate the impact of perturbation rate scaling from 0% to 100% with a step of 10%. We prioritize perturbing samples with higher weight, which may include FN examples. Our observa-



(a) AUC

(b) TPR@1%FPR

Figure 5: Influence of prompt text length.



(a) AUC

(c) AUC

(b) TPR@1%FPR

(d) TPR@1%FPR

Figure 6: Influence of reference dataset volume and perturbation rate.

tions reveal that both high and low perturbation rates result in relative poor performance. A high perturbation rate may make FP examples harder to learn, leading to a lower TPR@1%FPR, while a low perturbation rate may also be detrimental due to untreated FN samples. Consequently, we have opted for a final perturbation rate of 50%.

4.4 Ablation Study

We conduct an ablation study to evaluate the improvement in performance attributed to the two introduced stages. Specifically, we remove the weight-based data augmentation and perturbation on hight-weight data, and report the results in Table 3. The primary attack represents the Neighbor Attack. We observe that the weight-based data augmentation plays a more significant role in our framework since it is the key stage to capture and magnify the nonmember signals. Furthermore, the results also demonstrate the impact of perturbation on high-weight data for enhancing AUC performance.

4.5 Robustness of our DF-MIA

In this section, we study the robustness of DF-MIA against differential privacy techniques.

Lately, LLMs have faced numerous privacy threats such as model extraction attack (He et al. 2021), membership inference attack (Mattern et al. 2023; Shi et al. 2024), and data extraction attack (Carlini et al. 2021; Zhang, Wen, and Huang 2023), gaining attention from the research community. Among the defense techniques, DP-SGD (Abadi et al. 2016) which is based on differential privacy (DP) (Dwork et al. 2006), stands out as a mainstream privacy protection method. It offers privacy guarantees by limiting the impact of a single sample on model parameters through the addition of noise to the clipped gradients. A smaller privacy budget ϵ signifies a more stringent level of protection, but may potentially impact the model's downstream performance. We have fine-tuned GPT-neo-1.3B in a manner consistent with the approach described in the existing study by (Li et al. 2021) on the AGnews dataset enhanced with Neighbor Attack. Our experimental results in Table 4 demonstrate that DP-SGD can reduce privacy risks to a certain extent, but DF-MIA still represents a significant threat.

	AUC	TPR@1%FPR
primary attack	0.754	2.1%
w/o data augmentation	0.803	2.7%
w/o perturbation	0.844	9.6%
DF-MIA	0.879	9.8%

Table 3: Results from the ablation study conducted on GPT-Neo-1.3B using the AGnews dataset.

ϵ	15	30	60	+inf
AUC(primary)	0.598	0.663	0.692	0.754
AUC(Ours)	0.761	0.824	0.844	0.879
TPR@1%FPR(primary)	0.7%	1.2%	1.3%	2.1%
TPR@1%FPR(Ours)	0.8%	4.8%	4.8%	9.8%

Table 4: Results of DP-SGD.

5 Conclusion

In conclusion, this paper has addressed a critical gap in the Membership Inference Attack (MIA) field by introducing DF-MIA, a distribution-free framework tailored for fine-tuned Large Language Models (LLMs). Our framework leverages samples under test as a reference dataset, enhancing MIA's ability to differentiate between member and non-member data reliably. By intensifying the non-member signals through innovative data augmentation and perturbation approaches, DF-MIA not only addresses the unique challenges of fine-tuned environments but also reduces reliance on assumptions about training data distribution. The effectiveness of DF-MIA was validated across diverse LLMs with varying complexities, clearly outperforming traditional methods.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62306093 and

62376074), the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (Grant No. 2022B1212010005), and Shenzhen Science and Technology Program (Grant No. JSGGKQTD20221101115655027, SGDX20230116091244004, KJZD20231023095959002).

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308–318.

Aditya, H.; Chawla, S.; Dhingra, G.; Rai, P.; Sood, S.; Singh, T.; Wase, Z. M.; Bahga, A.; and Madisetti, V. K. 2024. Evaluating Privacy Leakage and Memorization Attacks on Large Language Models (LLMs) in Generative AI Applications. *Journal of Software Engineering and Applications*, 17(5): 421–447.

AI@Meta. 2024. Llama 3 Model Card.

Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J.; Pieler, M.; Prashanth, U. S.; Purohit, S.; Reynolds, L.; Tow, J.; Wang, B.; and Weinbach, S. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. In Fan, A.; Ilic, S.; Wolf, T.; and Gallé, M., eds., *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, 95–136. virtual+Dublin: Association for Computational Linguistics.

Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; and Tramer, F. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, 1897–1914. IEEE.

Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; and Zhang, C. 2023. Quantifying Memorization Across Neural Language Models. In *The Eleventh International Conference on Learning Representations*.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.-M.; Chen, W.; et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3): 220–235.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 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*, 265–284. Springer.

He, X.; Lyu, L.; Sun, L.; and Xu, Q. 2021. Model Extraction and Adversarial Transferability, Your BERT is Vulnerable! In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2006–2012. Online: Association for Computational Linguistics.

Ishihara, S. 2023. Training Data Extraction From Pre-trained Language Models: A Survey. In Ovalle, A.; Chang, K.-W.; Mehrabi, N.; Pruksachatkun, Y.; Galystan, A.; Dhamala, J.; Verma, A.; Cao, T.; Kumar, A.; and Gupta, R., eds., *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, 260–275. Toronto, Canada: Association for Computational Linguistics.

Li, M.; Wang, J.; Wang, J.; and Neel, S. 2023. MoPe: Model Perturbation based Privacy Attacks on Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13647–13660. Singapore: Association for Computational Linguistics.

Li, X.; Tramer, F.; Liang, P.; and Hashimoto, T. 2021. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*.

Liu, Y.; Zhao, Z.; Backes, M.; and Zhang, Y. 2022. Membership inference attacks by exploiting loss trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2085–2098.

Mattern, J.; Mireshghallah, F.; Jin, Z.; Schoelkopf, B.; Sachan, M.; and Berg-Kirkpatrick, T. 2023. Membership Inference Attacks against Language Models via Neighbourhood Comparison. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 11330–11343. Toronto, Canada: Association for Computational Linguistics.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2017. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*.

Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; and Gao, J. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

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

Mireshghallah, F.; Uniyal, A.; Wang, T.; Evans, D. K.; and Berg-Kirkpatrick, T. 2022b. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1816–1826.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1797–1807. Brussels, Belgium: Association for Computational Linguistics.

Nasr, M.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A. F.; Ippolito, D.; Choquette-Choo, C. A.; Wallace, E.; Tramèr, F.; and Lee, K. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Peris, C.; Dupuy, C.; Majmudar, J.; Parikh, R.; Smaili, S.; Zemel, R.; and Gupta, R. 2023. Privacy in the time of language models. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, 1291–1292.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Sablayrolles, A.; Douze, M.; Schmid, C.; Ollivier, Y.; and Jégou, H. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, 5558–5567. PMLR.

Shi, W.; Ajith, A.; Xia, M.; Huang, Y.; Liu, D.; Blevins, T.; Chen, D.; and Zettlemoyer, L. 2024. Detecting Pretraining Data from Large Language Models. In *The Twelfth International Conference on Learning Representations*.

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

Song, L.; and Mittal, P. 2021. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2615–2632.

Thangarasa, V.; Gupta, A.; Marshall, W.; Li, T.; Leong, K.; DeCoste, D.; Lie, S.; and Saxena, S. 2023. SPDF: Sparse Pre-training and Dense Fine-tuning for Large Language Models. In Evans, R. J.; and Shpitser, I., eds., *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, 2134–2146. PMLR.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Vaithilingam, P.; Zhang, T.; and Glassman, E. L. 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.

Wang, J. G.; Wang, J.; Li, M.; and Neel, S. 2024. Pandora's White-Box: Increased Training Data Leakage in Open LLMs. *arXiv preprint arXiv:2402.17012*.

Wang, Y. 2024. A Tutorial on the Pretrain-Finetune Paradigm for Natural Language Processing. *arXiv preprint arXiv:2403.02504*.

Xiao, L.; and Chen, X. 2023. Enhancing llm with evolutionary fine tuning for news summary generation. *arXiv preprint arXiv:2307.02839*.

Ye, J.; Maddi, A.; Murakonda, S. K.; Bindschaedler, V.; and Shokri, R. 2022. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 3093–3106.

Yeom, S.; Giacomelli, I.; Fredrikson, M.; and Jha, S. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, 268–282. IEEE.

Yetiştiren, B.; Özsoy, I.; Ayerdem, M.; and Tüzün, E. 2023. Evaluating the code quality of ai-assisted code generation tools: An empirical study on github copilot, amazon codewhisperer, and chatgpt. *arXiv preprint arXiv:2304.10778*.

Yu, W.; Pang, T.; Liu, Q.; Du, C.; Kang, B.; Huang, Y.; Lin, M.; and Yan, S. 2023. Bag of tricks for training data extraction from language models. In *International Conference on Machine Learning*, 40306–40320. PMLR.

Zeng, S.; Li, Y.; Ren, J.; Liu, Y.; Xu, H.; He, P.; Xing, Y.; Wang, S.; Tang, J.; and Yin, D. 2024. Exploring Memorization in Fine-tuned Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3917–3948. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, B.; Haddow, B.; and Birch, A. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, 41092–41110. PMLR.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhang, Z.; Wen, J.; and Huang, M. 2023. ETHICIST: Targeted Training Data Extraction Through Loss Smoothed Soft Prompting and Calibrated Confidence Estimation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12674–12687. Toronto, Canada: Association for Computational Linguistics.