Privacy Evaluation Benchmarks for NLP Models

Wei Huang[1*], **Yinggui Wang**[1*†], **CenChen**[2]
[1]Ant Group, China
[2]East China Normal University, China
hw378176@antgroup.com, wyinggui@gmail.com, cenchen@dase.ecnu.edu.cn

Abstract

By inducing privacy attacks on NLP models, attackers can obtain sensitive information such as training data and model parameters, etc. Although researchers have studied, in-depth, several kinds of attacks in NLP models, they are non-systematic analyses. It lacks a comprehensive understanding of the impact caused by the attacks. For example, we must consider which scenarios can apply to which attacks, what the common factors are that affect the performance of different attacks, the nature of the relationships between different attacks, and the influence of various datasets and models on the effectiveness of the attacks, etc. Therefore, we need a benchmark to holistically assess the privacy risks faced by NLP models. In this paper, we present a privacy attack and defense evaluation benchmark in the field of NLP, which includes the conventional/small models and large language models (LLMs). This benchmark supports a variety of models, datasets, and protocols, along with standardized modules for comprehensive evaluation of attacks and defense strategies. Based on the above framework, we present a study on the association between auxiliary data from different domains and the strength of privacy attacks. And we provide an improved attack method in this scenario with the help of Knowledge Distillation (KD). Furthermore, we propose a chained framework for privacy attacks. Allowing a practitioner to chain multiple attacks to achieve a higher-level attack objective. Based on this, we provide some defense and enhanced attack strategies. The code for reproducing the results can be found at https://github.com/user2311717757/nlp_doctor.

1 Introduction

In the past few decades, research in the field of NLP-based machine learning, especially deep learning, has achieved significant progress. However, the advancement of these applications has also led to serious security and privacy risks. In particular, inference attacks (Zhang et al., 2020; Mehnaz et al., 2022; He et al., 2022; Mireshghallah et al., 2022) enable attackers to deduce critical user privacy information such as training data and target model parameters. In general, current privacy attacks are studied under various threat models and experimental setups, but they are typically examined in isolation. It necessitates a comprehensive understanding of the risks these attacks pose, including the common factors influencing their performance, the scenarios where different inference attacks can be applied, the effectiveness of defenses, and the relationships between the attacks. Therefore, we need a benchmark to holistically assess the privacy risks faced by NLP models. To fill this gap, we conduct a comprehensive privacy risk assessment of NLP models targeting four representative inference attacks and have open-sourced an NLP privacy evaluation benchmark.

Attacks and Defenses for NLP models: In this paper, we focus on studying four representative privacy attacks on both large (Llama2 (Touvron et al., 2023), Qwen (Bai et al., 2023), and GPT2-xl (Radford et al., 2019)) and small (BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and GPT2-small (Radford et al., 2019)) NLP models: Membership Inference Attack (MIA, LMIA) (Shokri et al., 2016), Model Inversion Attack (MDIA, LMDIA) (Zhang et al., 2020), Attribute Inference Attack (AIA, LAIA) (Mehnaz et al., 2022), and Model Extraction Attack (MEA, LMEA) (He et al., 2022). Adding an "L" before the name indicates an attack targeting LLM. Besides the four attack methods, we integrate some comprehensive defense methods, including DP-SGD (Abadi et al., 2016), SELENA (Tang et al., 2022), and Texthide (Huang et al., 2020). Although there is an existing evaluation system for privacy attacks in the image do-

[*]These authors contributed equally to this work
[†]Corresponding author (wyinggui@gmail.com).

Findings of the Association for Computational Linguistics: EACL 2024, pages 2615–2636

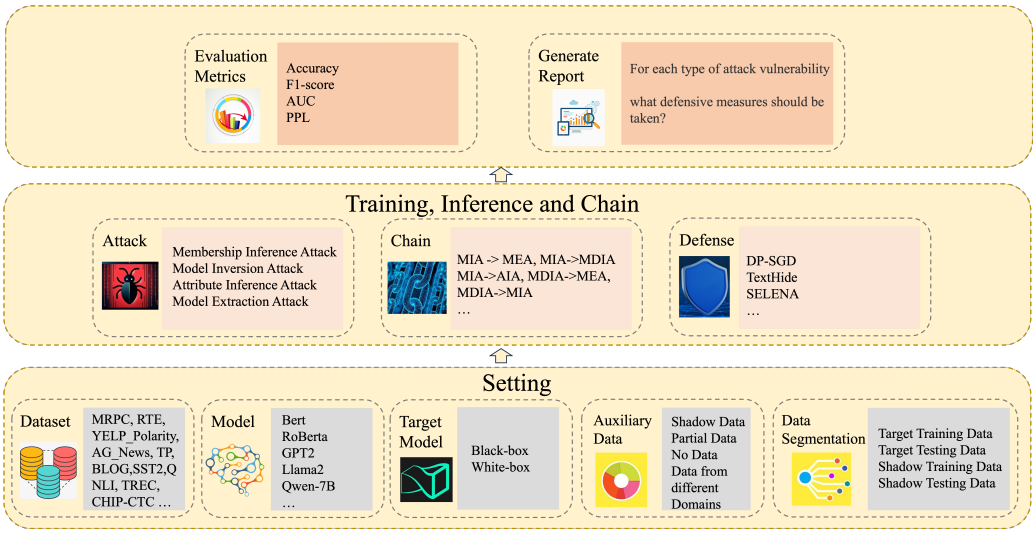

Figure 1: Overview of our privacy evaluation benchmark for NLP models.

main (Liu et al., 2022), our work differs from it in the following ways: 1) This paper focuses on NLP models, not the image-domain ones. 2) We introduce a privacy attack module for LLMs. 3) We explore the impact of auxiliary data from different domains on attacks. 4) We propose a chained framework for privacy attacks.

Attacks using Auxiliary Data from Different Domains: To fit the real scenarios (attackers may not obtain data with the same distribution as the auxiliary data), we conduct extensive experiments with data from different domains. From the experimental results (see Appendix E), we find that there are some cases where the attack performance is very low. Through our analysis, we conclude that the main reason for the poor performance of the attack is that the data distribution from different domains may be far from the original data distribution. We use KD to mitigate this issue.

A Chained Framework for Different Attacks: From an attack perspective, we hope to enable attackers to chain multiple attacks to achieve a higher-level attack objective. Secondly, we aim to explore whether one attack can influence the capability of another, in order to clarify the sequential relationship among different attacks, the inherent correlational characteristics among them, and the possibility of linking different attacks together. Based on the considerations mentioned above, we propose a chained framework for different attacks. From this, we provide some defense and enhanced attack methods: a defense method against MIA, a data-free MIA, and some strategies to improve the success rate of AIA, MDIA, and MEA.

The main contributions of this paper are as follows: In order to promote development and practical deployment of NLP privacy evaluation, we present a privacy evaluation benchmark for NLP models, which includes the conventional NLP models and LLMs. This paper experimentally analyzes the relationship between the data from different domains and attack performance. We use KD to address the problem of the low success rate of MIA under different domains, and also propose a chained framework for privacy attacks. It is mainly to enable a practitioner to chain multiple attacks to achieve a higher-level objective or explore the inherent correlational characteristics among attacks.

2 Overview of Privacy Evaluation Benchmarks

2.1 Implemented Components

Figure 1 provides an overview of the privacy evaluation benchmark. It primarily integrates modules for privacy attacks, defenses, and evaluations for various conventional NLP models and LLMs. The benchmark encompasses a wide range of threat models and usage scenarios, covering settings for target models in both black-box and white-box contexts. Additionally, it provides auxiliary data settings with shadow data, partial data, no data, and data from different domains. The benchmark also features data partitioning strategies for target training data, target testing data, shadow training data, and shadow testing data. In addition, we have integrated a chained framework that serves as an extension module for privacy attacks, aiming to explore higher attack objectives and the interrelationships between different attacks. Currently, our privacy evaluation benchmark supports 3 conven-

tional models and 3 LLMs, corresponding to 16 and 9 types of privacy attacks, respectively. In addition, it supports 14 types of chained connections and 3 defense mechanisms, as well as a wealth of evaluation strategies. Altogether, it supports 15 datasets from various tasks, including but not limited to classification and generation. We have provided a comprehensive description of the benchmark's usage process, with intricate details presented in Appendix A. Below, we will provide a detailed description of the settings for different threat models, the objectives of each attack, and how to use this benchmark.

2.2 Model Description

In this work, we focus on the classification (conventional models) and generation tasks (LLMs) in the NLP domain, which are among the most popular NLP applications. Typically, the goal of NLP classification tasks is to map data samples to a category. The output of an NLP classification model is a probability vector. And the output of an NLP generation model is text generated by the model based on the language patterns it has learned.

2.3 Threat Model

In different studies of privacy attacks, researchers have different restrictions on the knowledge that can be accessed by attackers, focusing on two aspects: access to the target model and auxiliary data.

Target Model: Referring to existing work, we can categorize the way an attacker accesses the model into two different setups: black box and white box. The former is denoted as B^{box}, which means that the attacker knows nothing about the internal structure of the target model. The latter is denoted as W^{box}, which means that the attacker has all the information about the target model, including model parameters, structure, and so on.

Auxiliary Data: In this paper, we classify the auxiliary data that an attacker can obtain into four cases: shadow data, partial data, no data, and auxiliary data from different domains. For the first one, it is denoted as D^{sha}, which means that the attacker has the same distributed data as the auxiliary data. The second one is written as D^{par}, which symbolizes that the adversary can get a part of the target data. For the third one, D^{no}, means that the attacker does not have any auxiliary data. For the fourth one, it is denoted as D^{diff}, which means that the attacker can only obtain data that has a different distribution from the training data.

Data Segmentation: To comply with the above assumption for the auxiliary data, we randomly divide the training data set into four non-overlapping sub-datasets of the same size as follows: Target Training Data, Target Testing Data, Shadow Training Data, and Shadow Testing Data. The first one is used to train the target model. The second one is used to evaluate the performance of attacks. The third is used to train the shadow model for attacks and is also used as a query dataset for various attacks. The last one is used in the training of a classification model for MIAs.

2.4 The Attack Objectives

Here we formally define the attack objectives under different attacks. The goal of a Membership Inference Attack is for the attacker to determine whether a target data sample was used to train the NLP model. The goal of a Model Inversion Attack is for the attacker to infer the training data of the target model when given access to the target model and auxiliary data, allowing an attacker to directly learn information about the training data. The objective of an Attribute Inference Attack is for the attacker to learn additional attribute information about the training data that is unrelated to the original task, such as gender, age, etc. This attack is used to explore unintentional information leakage. The goal of a Model Extraction Attack is to extract the parameters of the target model. Ideally, an attacker would have the capability to obtain a model with performance very similar to that of the target model, thereby achieving the effect of model reuse. As mentioned above, privacy attacks pose serious privacy threats; therefore, there is an urgent need for a benchmark to assess the privacy risks faced by the NLP model.

3 Privacy Attacks and Defenses for NLP Models

3.1 Privacy Attacks for NLP Models

This paper will focus on four privacy attacks against NLP small models: Membership Inference Attacks (MIA), Model Inversion Attacks (MDIA), Attribute Inference Attacks (AIA), and Model Extraction Attacks (MEA). Due to space limitations, we briefly describe the integrated attacks under various threat models in the main text, with detailed attack specifics provided in the appendices B. Attack methods are summarized in Table 10.

We have integrated eight types of MIAs for dif-

ferent threat models, five of which are black-box attacks (B^{box}/(D^{sha}, D^{par} or D^{diff}) (Shejwalkar et al., 2021; Salem et al., 2019; Song and Mittal, 2021)), which rely on thresholds or classification models to achieve the attack objectives. There are three white-box attacks (White Box/(D^{sha}, D^{par} or D^{diff}) (Nasr et al., 2019)), which mainly utilize gradients to distinguish between members and non-members. Detailed descriptions are included in the Appendix B.1 and B.5.

We have integrated three types of MDIAs focusing on white-box settings (W^{box}/(D^{no} (Fredrikson et al., 2015), D^{sha} or D^{diff}) (Zhang et al., 2022; Dathathri et al., 2020)). In our research, we have not yet discovered black-box MDIA on NLP classification models. One can imagine how difficult it would be for an attacker to infer the training data if they can only obtain logits or even just labels. Detailed descriptions are included in the Appendix B.2 and B.5.

For AIAs, we have implemented two types of attacks based on different threat models: black-box and white-box (B^{box} and W^{box}/D^{sha} (Lyu et al., 2021; Song and Shmatikov, 2020)). Both attacks rely on training attack models to ascertain attribute information. Detailed descriptions are included in the Appendix B.3.

Our benchmark also includes three types of black-box MEAs (B^{box}/(D^{sha}, D^{par} or D^{diff}) (Lyu et al., 2021; He et al., 2021)). Since the primary target of such attacks is the model parameters, assuming a white-box target model would negate the purpose of the attack. We primarily focus on models released as APIs. Detailed descriptions are included in the Appendix B.4 and B.5.

3.2 Solutions to Low Success Rate of MIA

From the experimental results (see Appendix E), we observe that in some cases membership inference performance is 0.500. We think that it is mainly because the shadow models trained on data from different domains cannot effectively summarize the membership states of the target model. We thus use KD (Hinton et al., 2015) to alleviate the above problem.

We know that KD can distill the knowledge contained in the teacher model into the student model. In other words, the student model can simulate the behavior of the teacher (target) model. The specific flow of this strategy is shown in Figure 2. Concretely, the target model is queried using training data from different domains to obtain logits, followed by training the student model using the data and labels mentioned above. Afterwards, the model is used as a shadow model for MIAs, and finally, the training and inference are performed according to the black-box MIA proposed in Appendix B.1.

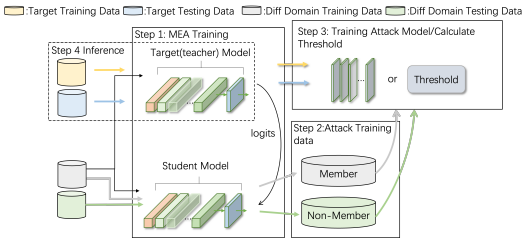

Figure 2: The flow chart for mitigating the low-performance issue of MIA under data from different domains with the help of KD.

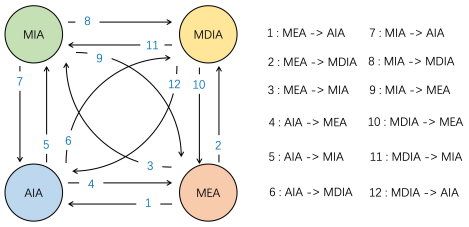

Figure 3: Framework for chaining different attacks.

3.3 Privacy Defenses for NLP Models

From the previous introduction, we have learned that privacy attacks can cause serious damage, so defenses against them are essential. Here, we select three classical or SOTA methods, namely, DP-SGD (Abadi et al., 2016), SELENA (Tang et al., 2022), and TextHide (Huang et al., 2020). Detailed descriptions are included in the Appendices B.6.

4 The Chained Framework

Most work on privacy attacks focused on a particular type of attack. Whether combining different attacks may yield different results is still an open problem. We propose a chained framework to enable a practitioner to chain multiple attacks to achieve a higher-level objective or explore the inherent correlational characteristics between attacks. Figure 3 shows the proposed chained framework.

MEA chained with AIA/MIA/MDIA: Figure 4 illustrates the flowchart of the method for chaining Model Extraction Attack (MEA) with other attacks. The initial step in this integration process involves utilizing shadow training data to query the target model, to train the extraction model. The subsequent steps are as follows.

MEA → AIA: The attacker uses shadow testing data to query the extraction model, which is then

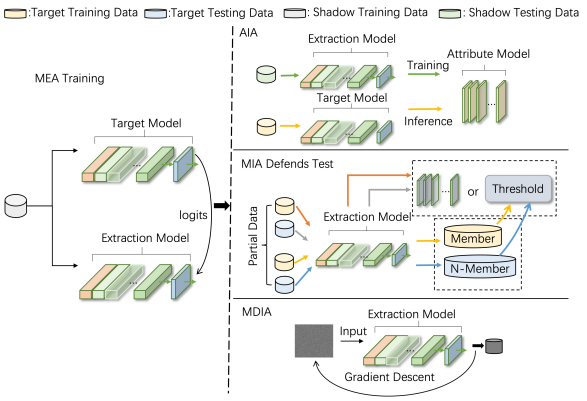

Figure 4: Flow chart of MEA chained with AIA/MIA/MDIA.

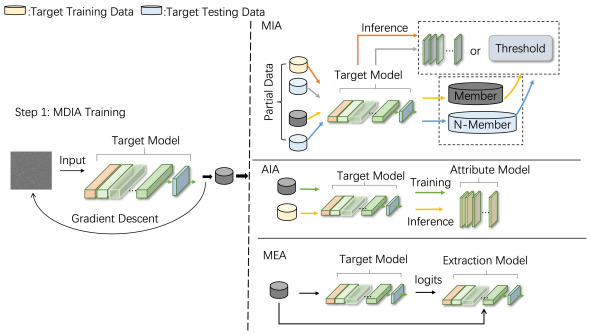

Figure 5: Flow chart of MDIA chained with MIA/AIA/MEA.

employed to train the attribute attack model. Finally, the attacker utilizes the target training data for inference. Since the attacker can access all parameters of the extracted model, they can deploy a white-box AIA method based on this model.

MEA → MIA: After the model owner publishes the extracted model, the attacker can proceed with the standard MIA procedure to conduct the attack test. Since MEAs leverage D^{sha} (shadow data) to obtain the extraction model, and D^{sha} does not belong to the target data, the extraction model can simulate the performance of the target model without inheriting the target model's training data. So, we consider that MEA can act as a defense strategy against MIA.

MEA → MDIA: This method employs the extracted model as a pseudo-target model for MDIAs to generate raw data. The purpose is to assess whether the extracted model retains any memory of the original training data.

MDIA chained with MIA/AIA/MEA: Figure 5 depicts the integration of MDIA with the remaining attacks. Given that the MDIA is a data-free reconstruction method, it is anticipated that the derived method can serve as a data generation strategy. This generated data can effectively supplant the assumption that the attacker possesses auxiliary data, thereby facilitating data-free attacks. The steps involved are **MDIA → MIA/AIA/MEA**, i.e., the attacker utilizes the data generated by the MDIA as auxiliary data to conduct the MIA, AIA, or MEA.

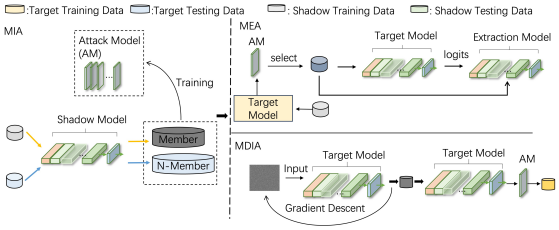

Figure 6: Flow chart of MIA chained with MEA/MDIA.

MIA chained with MEA/MDIA: Figure 6 illustrates chaining MIA with MEA or MDIA. MIA can identify whether a data point is part of the training set, thus it can serve as a data filter. The initial step involves training both the shadow model and the attack model. The subsequent steps include MIA → MEA and MIA → MDIA. In the **MIA → MEA** step, the shadow data is first processed by MIA to obtain the logits from the attack model, then the logits' member dimensions are ranked by score magnitude, and finally, the data is selected based on these scores proportionally. In **MIA → MDIA**, the attacker generates data using the MDIA, after which the generated data are subjected to the MIA testing process. The generated text is accepted only if the attack model classifies it as a member of the training data set.

Note that the attribute is just an implicit characteristic of a sample. It is irrelevant to the original task when training the target model, and it is just unconsciously remembered by the model. Obviously, in the NLP domain, an attacker who has determined a text attribute could do little to affect the performance of the other attacks.

5 Privack Attacks for LLM

In this study, we focus on four LLM privacy attacks, i.e., LMIA, LMDIA, LAIA, and LMEA. Detailed attack specifics are provided in the appendix C. Attack methods are summarized in Table 10.

Membership Inference Attack for LLM (LMIA): Based on different threat assumptions, we introduce three types of LMIAs. The first and second types require shadow data to train a reference model (B^{box}/D^{sha}) (Mattern et al., 2023; Mireshghallah et al., 2022). The third type (B^{box}/D^{diff}) (Fu et al., 2023) considers more realistic scenarios where the attacker can only obtain

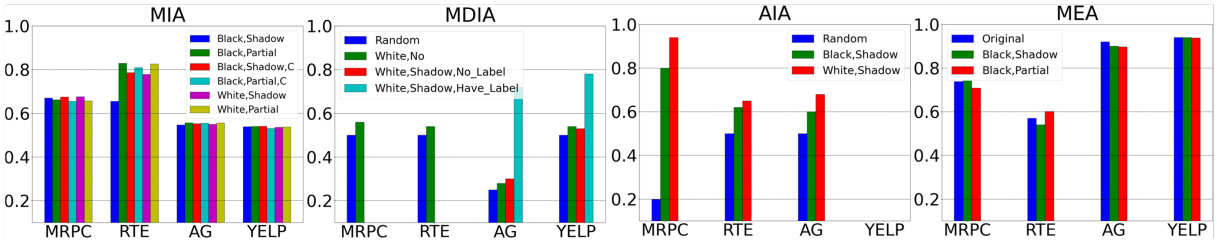

Figure 7: Attack accuracy of MIA/MDIA/AIA/MEA under different data for Bert.

data from different domains. Detailed descriptions are included in Appendix C.1.

Model Inversion Attack for LLM (LMDIA): We introduce two approaches for the black-box LMDIA(D^{no} or D^{diff}) (Carlini et al., 2021). This paper proposes a language-based LMDIA that employs text generation and Perplexity (PPL) sorting to extract batches of training text. Detailed descriptions are included in Appendix C.2.

Attribute Inference Attack for LLM (LAIA): We integrated two types of LAIA B^{box}/(D^{no} or D^{par}) (Staab et al., 2023; Lukas et al., 2023). This enables the LLM to extract attributes from the texts, including gender, address, occupation, and so on. For more detailed results, refer to Appendix C.3.

Model Extraction Attack for LLM(LMEA): This paper employs two attacks suitable for black-box scenarios (D^{sha} or D^{par}) (Tang et al., 2023; Gu et al., 2023). It mainly involves using an LLM to annotate unlabeled data. Detailed descriptions are included in Appendix C.4.

6 Results of Attacks and Defenses

In this section, we describe experimental data, the model, and the results of experiments on NLP privacy attacks and defenses.

Due to space constraints, we have presented only the most core experimental results in the main text, including privacy attack outcomes for both large-scale and traditional models as well as results from the chained framework. We have placed the performance of using data from different domains (see appendix E), the experimental results of KD (see appendix E), and the defense structures (see appendix F) in the appendix.

6.1 Datasets, Models, and Settings

In this paper, we select fifteen experimental datasets. They are: MRPC (Dolan and Brockett, 2005), RTE(Wang et al., 2019), YELP_Polarity (Zhang et al., 2015), AG_News (Zhang et al., 2015), TP (Hovy et al., 2015; Coavoux et al., 2018), BLOG (Schler et al., 2006; Coavoux et al., 2018), SST2 (Socher et al., 2013), QNLI (Wang et al., 2019), TREC (Li and Roth, 2002), CHIP-CTC (Zhang et al., 2021), KUAKE-QIC (Zhang et al., 2021), Wikitext-103 (Wiki) (Merity et al., 2016), ECHR (Chalkidis et al., 2019), Enron (Klimt and Yang, 2004) and PersonalReddit (PR) (Staab et al., 2023). The specifics of the data can be found in Appendix D.1.

To verify the effectiveness of different attack methods under different models, six commonly used NLP models, namely BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT2-small (subsequently abbreviated as GPT2) (Radford et al., 2019), Qwen-7B (Bai et al., 2023), Llama2 (Touvron et al., 2023)and GPT2-xl (Radford et al., 2019) are chosen for this study. We place the experimental setups and performance metrics in Appendix D.1.

6.2 Experimental Results for NLP Models

Due to space constraints, we only show results on Bert in the main text. The results for RoBert and GPT2 can be found in Appendix D.2.

Membership Inference Attack (MIA): From Figure 7, we can draw the following conclusions: 1) The success rate of the MIA is positively correlated with the degree of overfitting. The degree of model overfitting can be seen in Tables 12 and 13 in the appendix D. 2) White-box attacks' accuracy is better than black-box attacks', but the improvement is limited. 3) The accuracy of attacks in which the attacker has access to partial data is slightly higher than that of attacks using shadow data. 4) We find that the performance of the MIA and MEA are negatively correlated, a target model with higher MIA risks is less vulnerable to the MEA. For more details, refer to Appendix D.2.

Model Inversion Attack (MDIA): From Figure 7, we can draw the following conclusions: 1) The accuracy of MDIA is directly proportional to the complexity of the data. 2) When the auxiliary data is labeled, the performance of the attack is significantly better than that when it is unlabeled.

		Threat model	MRPC		RTE		AG_News		YELP		BLOG		TP	
			Ori	Acc	Ori	Acc	Ori	Acc	Ori	Acc	Ori	Acc	Ori	Acc
MEA	AIA	<Black,Shadow>	-	-	-	-	0.799	0.79	-	-	0.632	0.639	0.605	0.617
		<White,Shadow>	-	-	-	-	0.947	0.952	-	-	0.659	0.647	0.681	0.663
	MDIA	<Black,Shadow>	0.56	0.52	0.54	0.50	0.28	0.24	0.54	0.50	-	-	-	-
	MIA	<Black,Partial>	0.662	0.508	0.829	0.563	0.557	0.510	0.540	0.506	-	-	-	-
		<White,Partial>	0.657	0.506	0.826	0.547	0.556	0.515	0.538	0.504	-	-	-	-
MDIA	AIA	<Black,No>	-	-	-	-	0.799	0.540	-	-	0.632	0.549	0.605	0.517
		<White,No>	-	-	-	-	0.947	0.588	-	-	0.659	0.544	0.681	0.573
	MEA	<Black,No>	-	-	-	-	0.900	0.658	0.950	0.665	-	-	-	-
	MIA	<Black,No>	0.662	0.597	0.829	0.652	0.557	0.559	0.540	0.540	-	-	-	-
		<White,No>	0.657	0.593	0.714	0.714	0.556	0.540	0.538	0.545	-	-	-	-
MIA	MEA	<Black,Shadow>	0.748	0.755	0.542	0.523	0.900	0.883	0.950	0.944	-	-	-	-
		<Black,Partial>	0.711	0.714	0.603	0.625	0.898	0.860	0.947	0.931	-	-	-	-
	MDIA	<White,Shadow>	0.56	0.58	0.54	0.62	0.28	0.28	0.54	0.56	-	-	-	-
		<White,Shadow>	-	-	-	-	0.30	0.30	0.53	0.54	-	-	-	-

Table 1: Experimental results of the chained framework for the BERT model.

Methods	KUAKE-IR		CHIP-CTC		KUAKE-QTR	
	Llama2	Qwen	Llama2	Qwen	Llama2	Qwen
LiRA(Black Box/D^{sha})	0.541	0.535	0.524	0.519	0.557	0.542
LOSS Attack(Black Box/D^{sha})	0.607	0.624	0.572	0.558	0.682	0.671
SPV-MIA(Black Box/D^{diff})	0.688	0.676	0.604	0.583	0.761	0.689

Table 2: The AUC of membership inference attacks on LLM under different threat models.

Strategy	Text Generation Strategy		
	Top (D^{no})	Temp (D^{no})	Int (D^{diff})
Perplexity	5	3	19
Small	14	20	28
Medium	18	10	22
zlib	21	18	30
Window	11	12	20
Lowercase	24	10	30
Total Unique	93	73	149

Table 3: The number of memorized examples (out of 50 candidates) that we identify using each of the three text generation strategies and six membership inference techniques.

Model	MRPC	QNLI	RTE	SST2
Llama2	0.743/0.576	0.910/0.768	0.581/0.585	0.905/0.890
Qwen	0.743/0.777	0.910/0.768	0.581/0.621	0.905/0.894

Table 4: Results of the LMEA-G (D^{sha}). Original(BERT)/Acc.

Model	CHIP-CTC		KUAKE-QIC	
	Original	ACC	Original	ACC
Llama2	0.771	0.765	0.785	0.778
Qwen	0.782	0.770	0.791	0.790

Table 5: Results of the LMEA-I (D^{sha}, Lora).

For more details, refer to Appendix D.2. **Attribute Inference Attack (AIA):** From Figure 7, we can draw the following conclusions: 1) The success rate of attribute recognition for named entities will be higher than that for recognizing gender and age. 2) The performance of the white box is higher than that of the black box. For more details, refer to Appendix D.2.

Model Extraction Attack (MEA): From Figure 7, we can draw the following conclusions: 1) MEAs generally can achieve high accuracy. 2) The performance using partial data is generally lower than using shadow data. 3) For models with overfitting, the experimental results tend to perform a little better when the temperature hyperparameter is lower. For more details, refer to Appendix D.2.

Model	CHIP-CTC		KUAKE-QIC	
	Original	ACC	Original	ACC
Llama2	0.771	0.768	0.785	0.774
Qwen	0.782	0.780	0.791	0.790

Table 6: Attack Results of the LMEA-I (D^{par}, Lora).

	GPT2-Medium	GPT2-Large	GPT2-XL
ECHR	3.38%	17.37%	14.21%
Enron	7.66%	12.68%	15.55%

Table 7: The attack success rate of LAIA (Black Box/Partial Data) on all datasets.

6.3 Results of the Chained Framework

The data in Table 1 leads us to deduce the subsequent conclusions: 1) The extraction model provides a strong defense against MIAs. This implies that if the model owner employs the MEA and publishes the extracted model, it could serve as an effective defense method against MIAs. 2) When an attacker (black-box) conducts an AIA on the extracted model in a white-box context, the attack performance is superior to that in a black-box scenario. 3) Leveraging the attack capability of the MDIA, we demonstrate a data-free MIA, whose

Attr.	gender	location	married	age	education	occupation	pobp	income	Avg
Acc.	0.88	0.49	0.43	0.71	0.35	0.45	0.45	0.44	0.53

Table 8: The attack success rate of LAIA (Black Box/No Data) on all attributes in the PR dataset (Llama2-7B).

Attr.	gender	location	married	age	education	occupation	pobp	income	Avg
Acc.	0.88	0.46	0.50	0.68	0.38	0.49	0.50	0.46	0.54

Table 9: The attack success rate of LAIA (Black Box/No Data) on all attributes in the PR dataset (Llama2-13B).

accuracy is comparable to that of an attacker in possession of data with the same distribution. 4) We find that the MIA enhances the performance of the MEA on certain datasets (MRPC and RTE) when employed as a data filter. 5) Furthermore, on most datasets, data generated by the MDIA and filtered through the MIA can effectively increase the success rate of the MDIA. For more details, refer to Appendix G.

6.4 Experimental Results for LLM Attacks

Membership Inference Attack for LLM (LMIA): From Table 2, we can draw the following conclusions: 1) Existed MIAs designed for LMs (based on overfitting) cannot handle LLMs with large-scale parameters. However, methods based on memorization are suitable for LLMs. 2) The LMIA does not necessarily require the use of shadow data; satisfactory attack performance can also be achieved with data from different domains. For more detailed results, refer to Appendix H.

Model Inversion Attack for LLM (LMDIA): From Table 3, we can draw the following conclusions: 1) The LMDIA is more inclined to recover news headlines and log files. 2) The attacker's success rate using cross-domain data is higher than that of the no-data assumption. For more detailed results, refer to Appendix H.

Attribute Inference Attack for LLM (LAIA): From Table 9, 8 and 7, we can draw the following conclusions: 1) The capability of LAIA is correlated with the size of the model; the larger the model, the higher the success rate of the attribute inference attacks. 2) Age and gender are attributes that are easier to infer, whereas education and occupation are more difficult to infer. For more details, refer to Appendix H.

Model Extraction Attack for LLM (LMEA): From Table 4, 5 and 6, we can draw the following conclusions: 1) Compared to llama2, Qwen is more vulnerable to the LMEA attack, consequently revealing more of its knowledge. 2) Regardless of the threat model and fine-tuning strategy employed, attackers can easily transfer knowledge from the target model to the extraction model. 3) General LLMs possess a stronger capacity to resist LMEA compared to domain-specific LLMs. For more detailed results, refer to Appendix H.

7 Related Work

Privacy Attacks: Shokri et al. (Shokri et al., 2016) present the first MIA on ML models in the black-box case: it trains multiple shadow models to simulate the behavior of the target model and uses multiple classification models to distinguish members and non-members. Fredrikson et al. (Fredrikson et al., 2014) is the first to propose model inversion attacks on classification models. Subsequently, they use back-propagation of gradients to recover face information in (Fredrikson et al., 2015). Song et al. (Song and Shmatikov, 2020) propose the use of classification models to improve the performance of attribute inference attacks and expose the link between overfitting and attack performance. Tramèr et al. (Tramèr et al., 2016) propose the first model extraction attack against machine learning APIs. Jia et al. (Jia et al., 2019) observe that when the attack model is a binary classifier, it is vulnerable to adversarial examples.

8 Conclusion

In this paper, we develop a privacy attack and defense system for NLP models (the conventional/small models and LLMs). To be more realistic, we have done extensive experiments with auxiliary data from different domains. We further use KD to mitigate the poor performance of MIAs. On the other hand, we propose a chained framework to enable a practitioner to chain multiple attacks to achieve a higher-level objective. In real-world applications, our system can do a comprehensive privacy evaluation for NLP models to enable users to fully understand the extent of the model's leak privacy before it is deployed.

9 Limitations

In this paper, we introduce a benchmark for privacy evaluation in NLP, incorporating a wide array of methods tailored to different threat models. Throughout our research and implementation, we have uncovered several unexplored areas in NLP privacy attacks, including a shortage of black-box model inversion attacks on small NLP models and the insufficient accuracy of white-box model inversion attacks. We are committed to advancing the research in this field. Additionally, we propose a chain framework in this paper. While we have identified 14 types of connections, we believe there are many more potential connections between different attacks that could achieve higher attack objectives. To date, our research has focused on pairwise attack connections, but we intend to investigate more complex connections, including those involving three or more than three attacks. Although Large Language Models (LLMs) are in their nascent stages and evolving rapidly, research on privacy attacks for these models is not as developed as for smaller models. We plan to prioritize privacy attacks on LLMs in our future work, aiming to contribute to the development of the privacy community surrounding LLMs.

References

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*, pages 308–318.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*.

Maximin Coavoux, Shashi Narayan, and Shay B Cohen. 2018. Privacy-preserving neural representations of text. *arXiv preprint arXiv:1808.09408*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 17–32.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2023. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. *arXiv preprint arXiv:2311.06062*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

Xuanli He, Chen Chen, Lingjuan Lyu, and Qiongkai Xu. 2022. Extracted bert model leaks more information than you think! *arXiv preprint arXiv:2210.11735*.

Xuanli He, Lingjuan Lyu, Qiongkai Xu, and Lichao Sun. 2021. Model extraction and adversarial transferability, your bert is vulnerable! *arXiv preprint arXiv:2103.10013*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on World Wide Web*, pages 452–461.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yangsibo Huang, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora. 2020. Texthide: Tackling data privacy in language understanding tasks. *arXiv preprint arXiv:2010.06053*.

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*, pages 259–274.

Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. In *CEAS*, volume 45, pages 92–96.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. 2022. ML-Doctor: Holistic risk assessment of inference attacks against machine learning models. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4525–4542. USENIX Association.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. *arXiv preprint arXiv:2302.00539*.

Lingjuan Lyu, Xuanli He, Fangzhao Wu, and Lichao Sun. 2021. Killing two birds with one stone: Stealing model and inferring attribute from bert-based apis. *arXiv preprint arXiv:2105.10909*.

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

Shagufta Mehnaz, Sayanton V Dibbo, Roberta De Viti, Ehsanul Kabir, Björn B Brandenburg, Stefan Mangard, Ninghui Li, Elisa Bertino, Michael Backes, Emiliano De Cristofaro, et al. 2022. Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4579–4596.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David K 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*, pages 1816–1826.

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 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 739–753. IEEE.

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):9.

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 *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS)*.

Jonathan Schler, Moshe Koppel, Shlomo Engelson Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

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

R. Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2016. Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Congzheng Song and Vitaly Shmatikov. 2020. Overlearning reveals sensitive attributes. In *8th International Conference on Learning Representations, ICLR*.

Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security Symposium*, page 4.

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

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. 2022. Mitigating membership inference attacks by Self-Distillation through a novel ensemble architecture. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1433–1450. USENIX Association.

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

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction apis. In *USENIX security symposium*, volume 16, pages 601–618.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *In the Proceedings of ICLR*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Ruisi Zhang, Seira Hidano, and Farinaz Koushanfar. 2022. Text revealer: Private text reconstruction via model inversion attacks against transformers. *arXiv preprint arXiv:2209.10505*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. 2020. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 253–261.

A How to Use Benchmark

The following steps are required to use the privacy evaluation benchmark:

Build: Download our code repository from GitHub and prepare all the required environments.

Configure: Modify the configuration JSON file to specify settings, including the model, auxiliary data, attack type, defense method, target model, and attack model training hyperparameters (such as 'epoch' and 'lr', etc.). Descriptions of all hyperparameters can be found in their respective folders, as seen in the README.md located in each folder.

Load Model and Data: Users can use Load_model() and Load_aux_data() to handle the uploaded model and auxiliary data, and use tokenizer() to encode texts. We leverage Hugging Face's (Wolf et al., 2020) open-source libraries to load models and data.

Attack & Chain & Defense: Users can navigate to the relevant folder as specified by the attack and defense types in the configuration file and execute shell scripts to run the attack and defense mechanisms. To operate within the chained framework, preliminary attacks must be conducted under the designated attack and threat models before proceeding with the follow-up attacks.

Evaluate: User can use compute_metrics() in run_glue.py to obtain results and evaluations. We summarize the above steps in Figure 8. Before deploying the target model in the real world, model owners can assess the privacy threats their models may face from our benchmark.

B Privacy Attacks for NLP Models

This paper will focus on four privacy attacks against NLP small models: MIA, MDIA, AIA, and MEA. MIA is to infer whether the target sample is included in the training data. MDIA is to reconstruct the original training samples. AIA obtains information about attributes of the training data that are not relevant to the original task. MEA is used to reconstruct the model parameters or model functions.

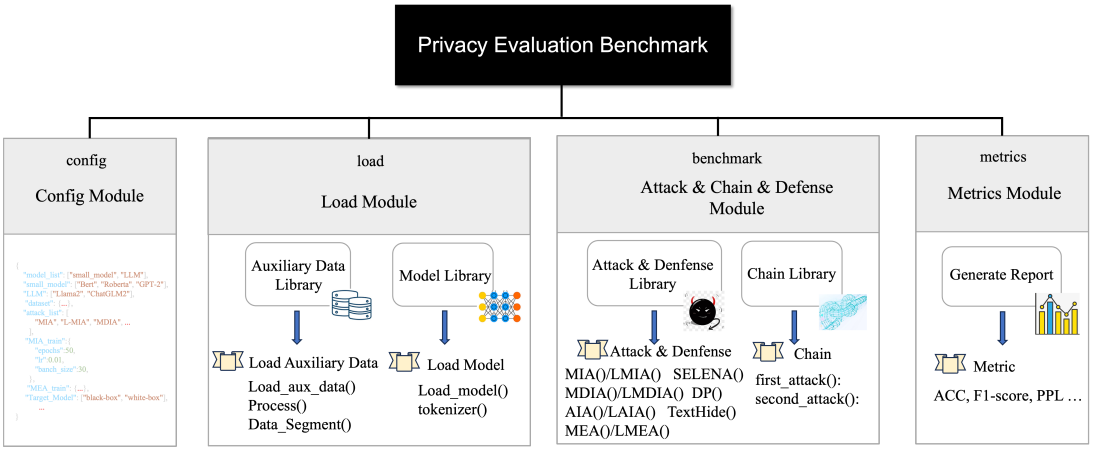

Figure 8: Workflow of benchmark

B.1 Membership Inference Attack (MIA)

The goal of MIA is to infer whether the target sample is included in the training data of the target model. Inferring the membership of the target sample can trigger a serious user privacy breach. For example, if the target model is trained using the user's bank transaction records, MIA will leak the user's financial and transaction details. Next, we describe how an attacker can execute a membership inference attack under different assumptions.

Black Box (Threshold or Classification Model based Methods)/(Shadow Data or Partial Data) (Shejwalkar et al., 2021; Salem et al., 2019; Song and Mittal, 2021): We introduce two approaches for the black-box MIAs. The former distinguishes members and non-members based on a threshold (Shejwalkar et al., 2021; Song and Mittal, 2021), and the latter trains a classification model for this purpose (Salem et al., 2019). In the first step, the attacker uses the shadow training data and shadow testing data to query the shadow model (the shadow model has been trained using the shadow training data) to get the attack feature (the first method uses losses obtained by comparing predictions with labels, the second method uses logits). In the second step, an adversary uses the attack feature obtained in the previous step to calculate the threshold for different labels (the loss of the member data is less than this threshold and vice versa) or to train the attack model (the label of the member data is 1 and the opposite is 0). In the third step, the target samples are fed to target model to obtain losses or logits, and finally the threshold or the attack model obtained in previous steps are used for judgment. In this scenario, the auxiliary data the attacker used are the shadow data or partial data. The difference is that the attacker in the case using the partial data does not need to train a shadow model, but directly queries the target model.

White Box (Classification Model based Methods)/(Shadow Data or Partial Data) (Nasr et al., 2019): The white-box MIA differs from the black-box one in that the attacker can access to all information of the model, so the attacker can train the attack model using gradients, outputs of the last activation layer, logits, classification losses and labels of the target model. The rest of steps are similar to that of the black-box attack.

B.2 Model Inversion Attack (MDIA)

The goal of MDIA is to reconstruct the original training samples, and there is some work in the field of image classification. It is difficult to implement this attack in NLP since the text is the serialized data. So we referring to the image MDIA methods.

White Box/(No Data) (Fredrikson et al., 2015): This approach assumes that the attacker can obtain all knowledge of the model and does not require any auxiliary data. First, a batch of noisy samples is generated and labels are randomly assigned to them. Then, the samples are fed into the target model to obtain losses, followed by updating this sample (here the sample serves as learnable parameters) using the gradient descent algorithm until the classification loss is less than a threshold set in advance or the number of iteration rounds is reached.

White Box/(Shadow Data) (Zhang et al., 2022; Dathathri et al., 2020): A model inversion attack in the field of NLP classification was first proposed by Ruisi Zhang et al. First, This work starts with an N-gram analysis using the shadow data to get the prompt texts for subsequent generation, called

templates (that paper considered the auxiliary data to be unlabeled data, The thesis also takes into account the scenario with labels.). Then it uses the shadow data to fine-tune GPT2. Finally it applies feedback from the target model to perturb the hidden state of the GPT2 model to generate texts.

B.3 Attribute Inference Attack

The goal of AIA is to infer attribute information of the target data. During the training process, the target model unintentionally learns additional information in addition to the classification information of the original task. In this case, the attacker can obtain extra attribute information (age, gender, name entity, etc.) of the corresponding sample.

Black Box and White Box/(shadow data) (Lyu et al., 2021; Song and Shmatikov, 2020): An attacker uses the shadow data to query the target model to obtain model outputs (logits for the black-box case and outputs of encoders for the white-boxe case) and uses attributes as labels to obtain the attack features. Next, the attack features are used to train the attack model. Finally, the target samples are fed into the target model, and the target model outputs predictions.

B.4 Model Extraction Attack

The goal of MEA is to reconstruct the parameters of the target model, so that the attacker can obtain an extracted model with a comparable performance to the target model. MEA is mainly based on the scenarios where the target model is accessed in the form of APIs.

Black Box/ (Shadow Data or Partial Data) (Lyu et al., 2021; He et al., 2021): The attacker queries the target model with the shadow data (unlabelled data) to obtain logits as soft labels, and then uses the shadow data and soft labels to train a function-extracted model.

B.5 Privacy Attacks using Auxiliary Data from Different Domains

To be more practical, we remove the assumption that the attacker has data from the same distribution, and instead that the adversary has data from different domains.

The attack strategy under data from different domains is very similar to previous privacy attacks. The difference is that the attacker uses an NLP dataset with different distributions from Target Training Data as an auxiliary dataset. We have conducted numerous experiments on MIAs, MIDAs, and MEAs (AIA are not taken into account here, because the attribute information varies across domains, which makes it difficult to obtain meaningful results).

Attack Strategy of Membership Inference: The attacker uses data from different domains to train the shadow model.

Attack Strategy of Model Inverse: The hacker under this attack uses data from different domains to query the target model to obtain the output of the embedding layer, and later uses this output as the initial sample for the attacker.

Attack Strategy of Model Extraction: In this scenario, we assume that the data obtained from different domains are unlabeled, and we query the target model for labeling to obtain soft labels.

B.6 Privacy Defenses for NLP Models

From the previous introduction, we have learned that privacy attacks can cause serious damage, so defenses against them are essential. Here, we select three classical or SOTA methods, namely, DP-SGD (Abadi et al., 2016), SELENA (Tang et al., 2022), and TextHide (Huang et al., 2020). specifically, DP-SGD adds Gaussian noise to the gradient during the model training. SELENA is a framework to train privacy-preserving models that induce similar behaviors on member and non-member inputs to mitigate MIA. Texthide is a text privacy protection technique and requires each participant to add a simple step to hide the representation of their text data with a one-time encryption.

C Attacks for LLM

In this study, we focus on four LLM privacy attacks, namely LMIA, LMDIA, LAIA, and LMEA.

C.1 Membership Inference Attack for LLM (LMIA):

Black Box/(Shaow Data) (Mireshghallah et al., 2022): We introduced a reference based attack, which adopts the pre-trained model as the reference model to calibrate the likelihood metric to infer membership (denoted as LiRA).

Black Box/(Data from Different Domains) (Mattern et al., 2023; Fu et al., 2023): In practical scenarios, assuming that the attacker has access to shadow data is a strong assumption. To better fit real-world situations, Mattern et al (Mattern et al., 2023), propose neighborhood attacks (Neighbour Attack), which compare model

	Target Model		Auxiliary Data				Attacker's Objective		
	B^{box}	W^{box}	D^{sha}	D^{par}	D^{no}	D^{diff}	Training Data	Attributes	Parameters
Conventional/Small Model (BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT2-small (Radford et al., 2019))									
<MIA, B^{box}, D^{sha}>	✓	-	✓	-	-	-	✓	-	-
<MIA, B^{box}, D^{par}>	✓	-	-	✓	-	-	✓	-	-
<MIA, B^{box}, D^{diff}>	✓	-	-	-	-	✓	✓	-	-
<MIA, W^{box}, D^{sha}>	-	✓	✓	-	-	-	✓	-	-
<MIA, W^{box}, D^{par}>	-	✓	-	✓	-	-	✓	-	-
<MIA, W^{box}, D^{diff}>	-	✓	-	-	-	✓	✓	-	-
<MDIA, W^{box}, D^{sha}>	-	✓	✓	-	-	-	✓	-	-
<MDIA, W^{box}, D^{no}>	-	✓	-	-	✓	-	✓	-	-
<MDIA, W^{box}, D^{diff}>	-	✓	-	-	-	✓	✓	-	-
<AIA, B^{box} , D^{sha}>	✓	-	✓	-	-	-	-	✓	-
<AIA, W^{box}, D^{sha}>	-	✓	✓	-	-	-	-	✓	-
<MEA, B^{box}, D^{sha}>	✓		✓	-	-	-	-	-	✓
<MEA, B^{box}, D^{par}>	-	✓	-	-	✓	-	-	-	✓
<MEA, B^{box}, D^{diff}>	-	✓	-	-	-	✓	-	-	✓
LLM (Llama2 (Touvron et al., 2023),Qwen (Bai et al., 2023),GPT2-xl (Radford et al., 2019))									
<LMIA, B^{box}, D^{sha}>	✓	-	✓	-	-	-	✓	-	-
<LMIA, B^{box}, D^{diff}>	✓	-	-	-	-	✓	✓	-	-
<LMDIA, B^{box}, D^{no}>	✓	-	-	-	✓	-	✓	-	-
<LMDIA, B^{box}, D^{diff}>	✓	-	-	-	-	✓	✓	-	-
<LAIA, B^{box}, D^{no}>	✓	-	-	-	✓	-	-	✓	-
<LAIA, B^{box}, D^{par}>	✓	-	-	✓	-	-	-	✓	-
<LMEA, B^{box}, D^{sha}>	✓	-	✓	-	-	-	-	-	✓
<LMEA, B^{box}, D^{par}>	✓	-	-	✓	-	-	-	-	✓

Table 10: An overview of the characteristics of different attacks and their corresponding threat models. ✓indicates the attack has this feature, – indicates the attack does not have this feature.

scores for a given sample to scores of synthetically generated neighbor texts and therefore eliminate the need for access to the training data distribution. For details, refer to Equation 1, where f_θ represents the target model, x denotes the target sample point, $\tilde{x}_i$ signifies the neighboring sample point (obtained by replacing words in x), and γ indicates the threshold value.

$$A_{f_\theta}(x) = \mathbb{1}[L(f_\theta, x) - \frac{\sum_{i=1}^{n} L(f_\theta, \tilde{x}_i)}{n} < \gamma] \quad (1)$$

Fu et al (Fu et al., 2023), 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, they introduce a more reliable membership signal, probabilistic variation, which is based on memorization rather than overfitting. Furthermore, they introduce a self-prompt approach, which constructs the dataset to fine-tune the reference model by prompting the target LLM itself. For details, refer to Equation 2, where $\tilde{p}_\theta(x)$ and $\tilde{p}_\phi(x)$ are probabilistic variations of the text record x measured on the target model θ and the reference model ϕ respectively.

$$A_{our}(x, \theta, \phi) = \mathbb{1}[\tilde{p}_\theta(x) - \tilde{p}_\phi(x) \leq \gamma] \quad (2)$$

C.2 Model Inversion Attack for LLM (LMDIA):

Black Box/(No Data/ Data from Different Domains) (Carlini et al., 2021): We introduce one approach for the black-box. This work first builds three datasets of 200,000 generated samples (each of which is 256 tokens long) using one of the following strategies:

- Top-n (Top): samples naively from the empty sequence.
- Temperature (Temp): increases diversity during sampling.
- Internet (Int): conditions the LM on Internet text.

We order each of these three datasets according to each of the six membership inference metrics:

- Perplexity: the perplexity of the largest GPT-2 model.
- Small: the ratio of log-perplexities of the largest GPT-2 model and the Small GPT-2 model.
- Medium: the ratio as above, but for the Medium GPT-2.
- zlib: the ratio of the (log) of the GPT-2 perplexity and the zlib entropy (as computed by compressing the text).
- Lowercase: the ratio of perplexities of the GPT-2 model on the original sample and on the lowercased sample.

- Window: the minimum perplexity of the largest GPT-2 model across any sliding window of 50 tokens.

For each of these 3×6 = 18 configurations, we select 100 samples from among the top 1000 samples.

C.3 Attribute Inference Attack for LLM (LAIA):

Black Box/(No Data or Partial Data) (Staab et al., 2023; Lukas et al., 2023): Recently, Staab et al. proposed a method for an LLM attribute inference attack. Specifically, Given t (text), A_1 (Adversary) first creates a prompt $P_{A_1}(t) = (S, P)$. For this, P_{A_1} is a function that takes in the text t and produces both a system prompt S and a prompt P which is given to the model M. While this formulation is general, for the rest of this work, they restrict the prompt P to $P = (PrefixF_{A_1}(t)Suffix)$ where F_{A_1} is a string formatting function. The specific form of the prompt can be seen in the original paper. The model M responds to this prompt with $M(P_{A_1}(t)) = (a_j, v_j)_{1 \leq j \leq k}$ (Where 'a' represents attribute, 'v' represents value.) the set of tuples it could infer from the text.

Lukas et al. proposed a PII (Personally Identifiable Information) reconstruction, they assume a more informed attacker, similar to that of membership inference, who has some knowledge about the dataset. For example, when an attacker wants to learn more PII about a user, they can form masked queries (e.g., "John Doe lives in [MASK], England") to the LM and attempt to reconstruct the missing PII.

C.4 Model Extraction Attack for LLM (LMEA):

Black Box/(Shaow Data or Partial Data): This paper employs two attacks suitable for black-box scenarios. The first approach denoted as LMEA-G (D^{sha} or D^{par}), targets generalized LLMs (Tang et al., 2023). It involves using an LLM to annotate unlabeled data, which is then used to train smaller models like BERT.

The second approach, LMEA-I (D^{sha} or D^{par}) (Gu et al., 2023), is tailored for industrial LLMs. We utilize strategies for parameter-efficient Tuning, such as LoRA (Hu et al., 2021) and P-Tuning v2 (Liu et al., 2021). This method relies on shadow or partial data to extract soft labels that are subsequently used to train the extraction model.

D Results of Attacks and Defenses

In this section, we describe experimental data, the model, and the results of experiments on NLP privacy attacks and defenses.

D.1 Datasets, Models, and Settings

In this paper, we selected fifteen experimental datasets. They are: MRPC (Dolan and Brockett, 2005), RTE(Wang et al., 2019), YELP_Polarity (Zhang et al., 2015), AG_News (Zhang et al., 2015), TP (Coavoux et al., 2018), BLOG (Schler et al., 2006), SST2 (Socher et al., 2013), QNLI (Wang et al., 2019), TREC (Li and Roth, 2002), CHIP-CTC (Zhang et al., 2021), KUAKE-QIC (Zhang et al., 2021), Wikitext-103 (Wiki) (Merity et al., 2016), ECHR (Chalkidis et al., 2019), Enron (Klimt and Yang, 2004) and PersonalReddit (PR) (Staab et al., 2023) (CHIP-CTC and KUAKE-QIC are used for LMEA experiments. ECHR, Enrom and PR are used for LAIA experiments). Because attribute inference attacks require data to have attribute labels, we introduced five additional datasets in this attack (TP, BLOG, ECHR, Enron and PR), and used the remaining datasets as supplementary data in experiments across different domains. The specifics of the data can be found in 11.

To verify the effectiveness of different attack methods under different models, six commonly used NLP models, namely BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT2-small (subsequently abbreviated as GPT2) (Radford et al., 2019), Qwen-7B (Zeng et al., 2023), Llama2 (Touvron et al., 2023) and GPT2-xl (Radford et al., 2019) were chosen for this study. The experimental setup can be seen in the description below, and the performance of the target model on each dataset is shown in Table 12 and 13.

Performance Metrics We use four metrics to measure Attack Performance and Defense Performance, namely Accuracy (the proportion of correctly predicted samples to the total number of samples, with higher values indicating better attack performance), F1-score (a composite of precision and recall, with higher values indicating better attack performance), PPL (Perplexity, used to reflect the fluency of the text, with lower values indicating better model counter-attack effectiveness), and AUC (Area Under the ROC Curve, is a metric used to measure the performance of binary classification models. The closer the AUC value is to 1, the bet-

Dataset	#Label	Attribute	#Train	#Dev	Task	Source
MRPC	2	\	3668	408	paraphrase	news
QNLI	2	\	103k	5K	QA/NLI	Wikipedia
RTE	2	\	2490	277	NLI	news/Wikipedia
AG_News	4	\	120K	7.6K	Topic Classification	news
SST2	2	\	67K	872	sentiment	movie reviews
TREC	6	\	5000	452	Topic Classification	QA
YELP_Polairty	2	\	560K	38K	sentiment	movie reviews
TP	5	age, gender	24K	2767	sentiment	Trustpilot Sentiment dataset
AG_News(AIA)	4	entity	13K	1457	Topic Classification	news
BLOG	10	age, gender	7985	887	Topic Classification	blog authorship corpus
CHIP-CTC	44	\	22k	7k	Medical	Clinical trial
KUAKE-QTR	11	\	6931	2K	Medical	Medical search
KUAKE-IR	2	\	5000	600	Medical	Medical search
ECHR	1	name	7.1k	1.38k	Legal	law cases dealt
Enron	4	e-mail, work address..	31.7k	2K	e-mail	e-mail
PersonalReddit	8	age, education..	1184	-	Reddit	Reddit profiles

Table 11: Dataset introduction

	BERT	RoBERTa	GPT2
MRPC	0.9640/0.7426	0.8974/0.7966	0.8337/0.7353
RTE	0.9775/0.5812	0.9389/0.6173	0.9695/0.5993
AG_News	0.9897/0.9243	0.9838/0.9353	0.9872/0.9318
YELP	0.9984/0.9498	0.9975/0.9603	0.9977/0.9585

Table 12: Performance of the target model on training/testing sets on each dataset.

	AG(AIA)	BLOG	TP
BERT	0.7934	0.9391	0.8652
RoBERTa	0.8058	0.9481	0.8699
GPT2	0.7907	0.9200	0.8771

Table 13: Inference accuracy of different models using different datasets.

ter the model is at distinguishing between positive and negative classes. An AUC value of 0.5 indicates that the model has no discriminative power, equivalent to random guessing.).

Target model training. In this paper, we chose to use bert-base-uncased as the BERT model, for the MRPC dataset we used a learning rate of 2e-5 and batch_size of 32. We completed all experiments on the open-source framework transformers.

Membership Inference Attack (Settings). We train the shadow model in the same way in different membership inference attacks, where we ensure that the architecture and training process of the shadow model and the hyperparameters chosen are consistent with the target model. In the approach of the classification model as an attack model, we construct different linear layers for different information. In this case, one linear layer was used for the loss values, and 3 layers were used for the logits, for the gradient information we chose to go through one CNN layer and then pick up 2 linear layers, and 2 linear layers were used for the labels. Finally, we combined the outputs of the different linear layers and fed them to the four linear layers. We use RELU as the activation function for the attack model, batch_size is set to 32, Adam as the optimizer, cross-entropy as the loss function and the learning rate is set to 1e-5.

Model Inversion Attack (Settings). In this paper, there are two different model inversion attacks. For the first model inversion attack, we set alpha as 50, beta as 20, gama as 0.001, learning_rate as 0.1, and set the number of recovered embeddings as 100. For the second model inversion attack, n was set to [4,7) and [6,8) for AG_News and YELP_Polarity, respectively, when doing the n-gram analysis. For training, we set kl_loss to 0, window_length to 0, step to 0.004, and set the number of repetitions to 10. In the evaluation phase, we selected the evaluation models Distil-Bert, Distil-RoBerta, and Distil-GPT2 and followed the transformers' script for training.

Attribute Inference Attack (Settings). For the attribute inference attack, we used 4 linear layers in training the attack model and set the learning rate as 1e-4, the optimizer as Adam, the batch_size as 32, and the loss function as the cross-entropy loss function.

Model Extraction Attack (Settings). When training the extraction model in this paper, the logits obtained by querying the target model with shadow data are used as soft labels. We ensure that the architecture and training process of the extraction model is consistent with the target model.

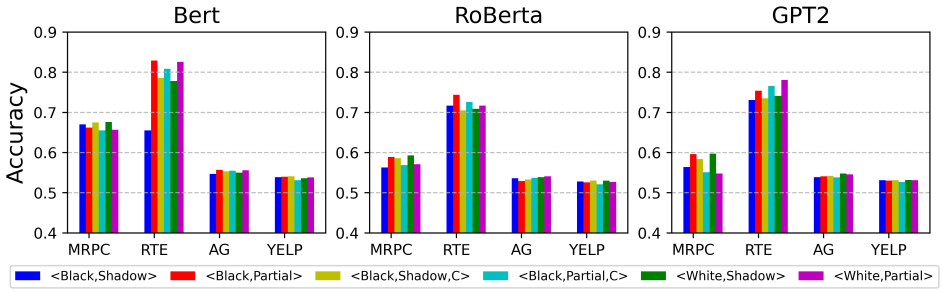

Figure 9: Attack accuracy of MIA under different data and model architectures.

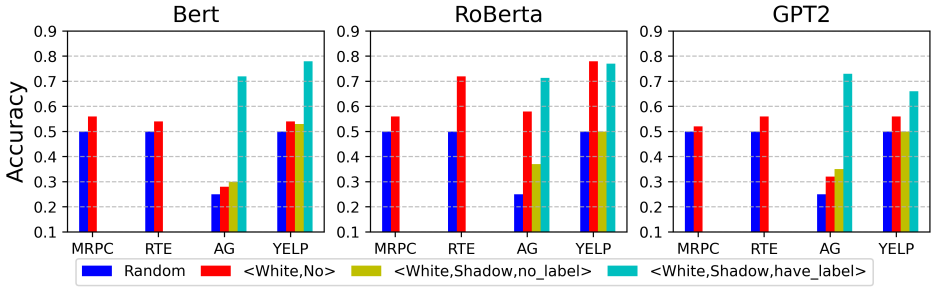

Figure 10: Attack accuracy of MDIA under different data and model architectures.

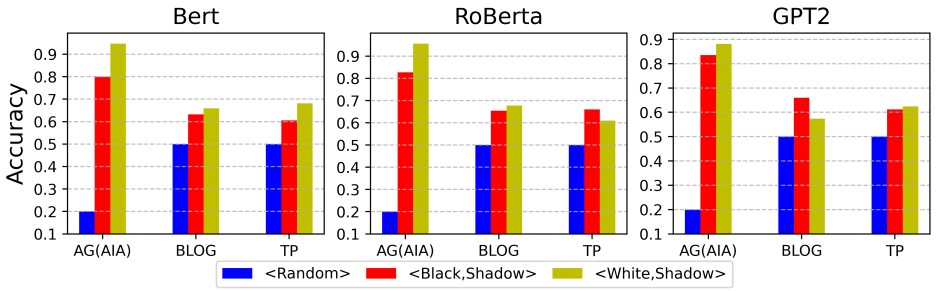

Figure 11: Attack accuracy of AIA under different data and model architectures.

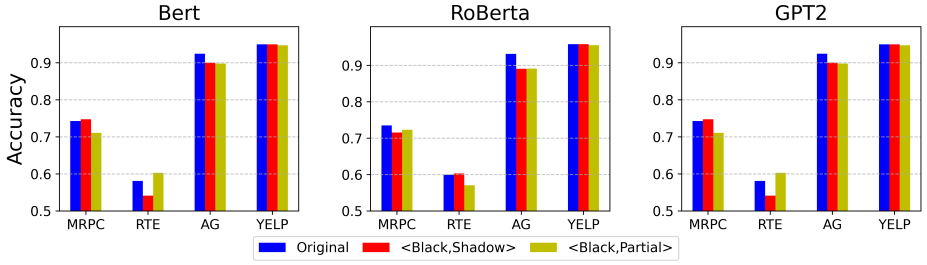

Figure 12: Attack accuracy of MEA under different data and model architectures.

D.2 Privacy Attack Performance

Membership Inference Attack (MIA): In Figure 9, we report the accuracy of MIA. We observe that the attack achieves high accuracy on both MRPC and RTE, for example, the BERT model trained on RTE dataset has a success rate of 0.778 for the attacker under the white-box assumption(shadow data). On the one hand, the attack performance on AG_News and YELP_Polarity is relatively low, mainly because these two datasets have better generalization performance, i.e. the model is less overfitted.

As we can see from Figure 9, when the attacker obtains the target model as white box, the success rate of attack is better than black box in most cases, although the difference is not substantial. Secondly, when the attacker obtains partial data, the attack performance is slightly better than shadow data, but also remains at the same level. We also verified that using the threshold approach or classification model approach had little effect on attack performance.

Model Inversion Attack (MDIA): The accuracy of MDIA can be seen in Figure 10. For the second

	AG(AIA)	BLOG	TP
<BERT,Black,Shadow>	0.4212	0.544	0.6086
<BERT,White,Shadow>	0.8850	0.6915	0.7062
<RoBERTa,Black,Shadow>	0.5258	0.7049	0.6485
<RoBERTa,White,Shadow>	0.9099	0.7171	0.6125
<GPT2,Black,Shadow>	0.5659	0.6955	0.6015
<GPT2,White,Shadow>	0.7259	0.4952	0.6170

Table 14: F1-scores of AIA.

Temprature	MRPC	RTE	AG_News	YELP
0.0	0.7353	0.5668	0.8970	0.9485
0.5	0.7475	0.5415	0.9004	0.9499
1.0	0.7426	0.5343	0.9004	0.9503
5.0	0.7426	0.5343	0.8987	0.9512

Table 15: The success rate of model extraction attacks at different temperatures for the BERT model using the shadow data.

attack, we only report performance on AG_News and YELP_Polarity, as the method can only recover single-sentence data. From Figure 10 we can see that BERT and GPT2 are not effective compared to RoBERTa in the MDIA (D^{no}). the success rate of the former two on RTE is around 0.55 compared to 0.72 for the latter, mainly because RoBERTa model learns more information about the data. On the second point, the performance of the attack is significantly better when the auxiliary data is labeled than when it is unlabeled. For example, when the target model is BERT trained on the YELP_Polarity dataset, the attack performance is 0.78 in the labeled scenario, compared to 0.53 in the unlabeled scenario.

Attribute Inference Attack (AIA): We measured the performance of AIA, and present the results in Figure 11. From figure, we can see that the success rate of attack is higher on AG_News (AIA) and relatively lower on BLOG and TP. One reason for this may be that the former's attribute is name entity, while the latter two were gender and age, and name entity is more likely to be learned by model. Secondly, we also observed that the white box's performance is higher than the black box, suggesting that the representation of the model contains more information about attributes. We also report metrics for the F1-score, which can be found in Table 14.

Model Extraction Attack (MEA): We show the performance of MEA in Figure 12 (taking temperature to be 0.5), and in general the attack achieves good results, for example, BERT model trained on YELP data achieves an accuracy of 0.950 (shadow data). Secondly, we can also see that the performance of partial data is generally lower than shadow data, one reason for this may be that the vector obtained by querying the target model with partial data has a lower entropy, which contains less information for the attacker to extract. We also report the impact of different temperatures on the attack, which can be seen in Table 15 and 16.

Temprature	MRPC	RTE	AG_News	YELP
0.0	0.7181	0.6029	0.8946	0.9466
0.5	0.7108	0.6029	0.8980	0.9473
1.0	0.7132	0.5848	0.8983	0.9464
5.0	0.7034	0.5487	0.8982	0.9478

Table 16: The success rate of model extraction attacks at different temperatures for the BERT model using the partial data).

Tables 15 and 16 show the success rate of the models in extracting attacks at different temperatures (BERT model), from which it can be seen that for models with suspected overfitting using the MRPC and RTE datasets, the experimental results tend to perform a little better when the temperature is lower. This suggests that they tend to prefer hard labels. In contrast, models with better generalization using the AG_news and Yelp_polarity datasets may tend to prefer labels with a higher temperature (softer). In general, the results of the experiments on different temperatures do not differ significantly.

E Performance using Auxiliary Data from Different Domains

Privacy Attacks: In sub-figures 1-5 of Figure 14, we show the attack performance of MIA, MDIA, and MEA for data from different domain scenarios. For both MIA and MEA attacks, we can obtain desirable results in most cases, even better than the case adopting the same data distribution hypothesis, such as in the black-box scenario when the target data is RTE and the data from different domains is MRPC (see the 1st subfigure). However, in some cases, the attack success rate is close to random guesses, e.g., for a black-box MIA attack, the attack success rate is 0.5 when the target data is MRPC and the data from different domains is SST2 (see the 1st subfigure). Another example is MEA, in which the target data is TREC and the data from different domains is SST2, the attack performance is 0.45 (see the 4th subfigure). For MDIA, the performance of attacks with the data from different domains is largely better than that of the first model inversion attack. This is also a

	MRPC	RTE	AG_News	YELP
Original	0.7426	0.5812	0.9243	0.9498
DP-SGD(ϵ = 5)	0.6838	0.4729	0.8925	0.8987
DP-SGD(ϵ = 15)	0.7011	0.5050	0.8972	0.9040
SELENA	0.7034	0.5740	0.9138	0.9533
TextHide	0.7230	0.5126	0.9067	0.9214

Table 17: Performance of the target model using defense methods.

	AG(AIA)	BLOG	TP
Orignal	0.793	0.939	0.865
DP-SGD(ϵ = 5)	0.741	0.882	0.834
DP-SGD(ϵ = 15)	0.745	0.891	0.840
SELENA	0.785	0.918	0.866
TextHide	0.764	0.906	0.795

Table 18: Performance of the target model using different defense methods.

reasonable observation since we obtain additional auxiliary data.

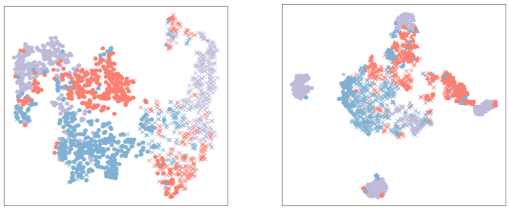

Figure 13: t-SNE plots for features obtained by querying the target model with data from different domains.

Cause Analysis: To analyze the problem of lower attack performance with data from different domains, we plotted the t-SNE plots for features obtained by querying the target model with data from different domains, as shown in Figure 13. In the first plot, the target data are SST2 (purple), and the data from different domains are AG_News (red) and TREC (blue) respectively. The graph shows that the representation distribution is closer, so it will work better. However, it can be seen in the second graph that the distribution of representation of the data from different domains is far from the target data (the target data is TREC (purple) and the data from different domains is AG_News (red) and SST2 (blue)), so this could explain why in some cases the attacker's success rate is lower.

Experimental Results of Improved Methods: Subfigure 6 in Figure 14 illustrates the results of experiments that mitigate the issue of the low success rate for membership inference under different domains. Compared with Subfigure 1, it can be seen that our proposed method of using the student model as a shadow model has great results. Among them, this method has significant improvement in the performance of membership inference attacks with a success rate close to 0.5. For example, when the target data is RTE and the data from different domains is QNLI, the success rate can reach 0.7.

F Defense Performance

In this offensive and defensive system, we integrate three defensive methods, namely DP-SGD, SELENA, and TextHide. We report the effect on the BERT model. Table 17 shows the performance of the target model after it has been protected by defense methods (a table of attribute inference attack can be found in Table 18), where we can see that DP-SGD has a larger impact on performance.

DP-SGD: In Tables 19 and 22, we report the effectiveness of DP-SGD's defense towards the four attacks. Overall, DP-SGD offers a more significant defense against the MIA, success rate of the attack is close to random guesses on the vast majority of the dataset. For example, on MRPC (black box/shadow data), DP-SGD has a defense capability of 0.492/0.504 against MIA, compared to an original attack performance is 0.670. For the other three attacks, the method is not significantly defensive but is still effective.

SELENA: Tables 20 and 22 show the effectiveness of SELENA's defenses, where the method is more effective against MIA, particularly MRPC and YELP. However, the defense against the other three attacks was poor, especially the AIA. For example, on the BLOG dataset, the original attack accuracy was 0.659, while SELENA performed 0.673. One possible reason for this is that the method does an aggregation operation on the output of multiple models, which allows the defense model to have more knowledge. Overall the method is not as good as DP-SGD at defending but has better identification accuracy than DP-SGD.

Texthide: In Tables 21 and 22, we show the defensive capabilities of TextHide, which is effective mainly against MIA and MEA, and poorly against MDIA and AIA. For example, on RTE (black box/shadow data), the defense ability for the first two are 0.503 and 0.509 respectively, while the original attack performance is 0.655 and 0.542. As we can see, the defensive effectiveness of this method is comparable to SELENA, but of course,

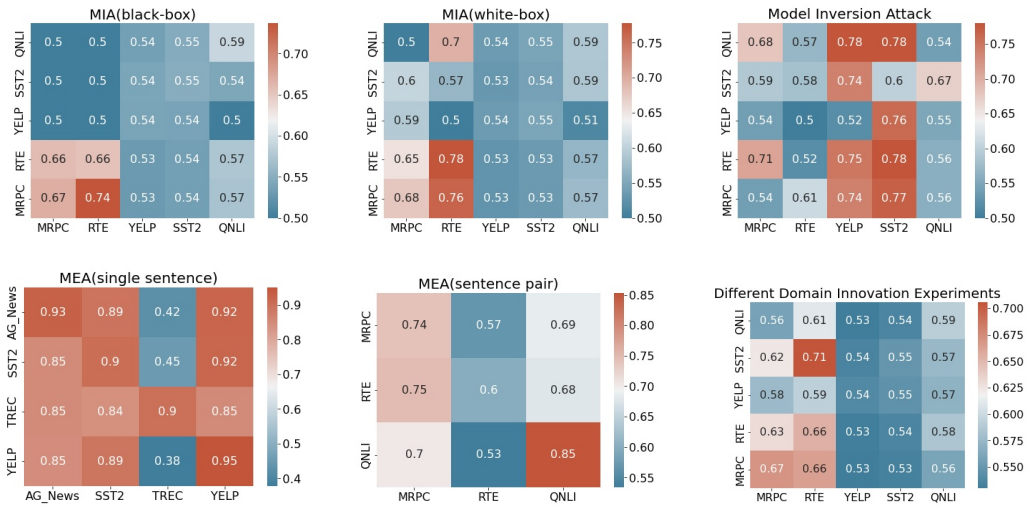

Figure 14: The experimental results show the performance of the BERT model on data from different domains. The horizontal axis represents the target training data, and the vertical axis represents data from different domains.

		MRPC		RTE		AG_News		YELP	
		Original	ϵ = 5/15	Original	ϵ = 5/15	Original	ϵ = 5/15	Original	ϵ = 5/15
MIA	<Black,Shadow>	0.670	0.492/0.504	0.655	0.504/0.513	0.547	0.509/0.508	0.539	0.501/0.497
	<Black,Partial>	0.662	0.493/0.509	0.829	0.513/0.548	0.557	0.506/0.509	0.54	0.500/0.501
	<White,Shadow>	0.676	0.495/0.498	0.778	0.501/0.515	0.550	0.508/0.505	0.536	0.502/0.506
	<White,Partial>	0.657	0.498/0.500	0.826	0.504/0.504	0.556	0.500/0.501	0.538	0.495/0.498
MDIA	<White,No>	0.56	0.54/0.50	0.54	0.50/0.50	0.28	0.26/0.26	0.54	0.50/0.52
	<White,Shadow>	-	-	-	-	0.30/0.72	0.3/0.69	0.53/0.78	0.46/0.70
MEA	<Black,Shadow>	0.748	0.684/0.699	0.542	0.473/0.484	0.9	0.883/0.888	0.95	0.908/0.913
	<Black,Partial>	0.711	0.683/0.699	0.603	0.473/477	0.898	0.881/0.884	0.947	0.910/0.913

Table 19: Experimental results of defense methods against MIA, MDIA, and MEA using DP-SGD for the BERT model.

		MRPC		RTE		AG_News		YELP	
		Original	Acc	Original	Acc	Original	Acc	Original	Acc
MIA	<Black,Shadow>	0.670	0.531	0.655	0.555	0.547	0.501	0.539	0.514
	<Black,Partial>	0.662	0.500	0.829	0.608	0.557	0.498	0.540	0.511
	<White,Shadow>	0.676	0.528	0.778	0.557	0.550	0.504	0.536	0.507
	<White,Partial>	0.657	0.497	0.826	0.605	0.556	0.505	0.538	0.511
MDIA	<White,No>	0.56	0.58	0.54	0.50	0.28	0.26	0.54	0.52
	<White,Shadow>	-	-	-	-	0.30/0.72	0.28/0.71	0.53/0.78	0.50/0.68
MEA	<Black,Shadow>	0.748	0.703	0.542	0.592	0.900	0.912	0.950	0.952
	<Black,Partial>	0.711	0.704	0.603	0.545	0.898	0.909	0.947	0.951

Table 20: Experimental results of defense methods against MIA, MDIA, and MEA using SELENA for the BERT model.

		MRPC		RTE		AG_News		YELP	
		Original	Acc	Original	Acc	Original	Acc	Original	Acc
MIA	<Black,Shadow>	0.670	0.564	0.655	0.503	0.547	0.523	0.539	0.507
	<Black,Partial>	0.662	0.570	0.829	0.510	0.557	0.527	0.540	0.506
	<White,Shadow>	0.676	0.553	0.778	0.503	0.550	0.526	0.536	0.508
	<White,Partial>	0.657	0.558	0.826	0.523	0.556	0.529	0.538	0.503
MDIA	<White,No>	0.56	0.56	0.54	0.50	0.28	0.28	0.54	0.50
	<White,Shadow>	-	-	-	-	0.30/0.72	0.26/0.73	0.53/0.78	0.44/0.72
MEA	<Black,Shadow>	0.748	0.689	0.542	0.509	0.9	0.892	0.95	0.934
	<Black,Partial>	0.711	0.683	0.603	0.524	0.898	0.893	0.947	0.936

Table 21: Experimental results of defense methods against MIA, MDIA, and MEA using TextHide for the BERT model.

it is still better than SELENA's method on MDIA, AIA, and MEA. Overall the method has a good balance between accuracy and defensive ability.

		AG_News(AIA)		BLOG		TP	
		Original	Acc	Original	Acc	Original	Acc
DP-SGD	<Black,Shadow>	0.800	0.805/0.805	0.632	0.652/0.649	0.605	0.618/0.617
	<White,Shadow>	0.947	0.945/0.947	0.659	0.649/0.694	0.681	0.617/0.682
SELENA	<Black,Shadow>	0.800	0.801	0.632	0.637	0.605	0.610
	<White,Shadow>	0.947	0.947	0.659	0.673	0.681	0.690
TextHide	<Black,Shadow>	0.800	0.811	0.632	0.621	0.605	0.619
	<White,Shadow>	0.947	0.923	0.659	0.639	0.681	0.649

Table 22: Experiment results of defense methods towards AIA for the BERT model. For DP-SGD, we report results with $\epsilon = 5, 15$.

G Results of the Chained Framework

Table 1 describes the experimental results of the chained framework. It can be seen that the extraction model provides a strong defense against membership inference attacks, particularly for the YELP dataset, where the attack performance is nearly 0.5. This implies that if the model owner employs MEA and publishes the extracted model, it could serve as an effective defense method against MIA. When an attacker conducts an AIA on the extracted model in a white-box context, the attack performance is superior to that in a black-box scenario. Take the TP dataset as an example: the attack success rate can be improved by approximately 6 percentage points. This is primarily because, after the attacker gains access to the function-extracted model using a black-box method, they can employ white-box knowledge to intensify the attack.

From Table 1, we can see that conducting a membership inference attack under the no-data condition is also feasible and yields a comparable attack performance to that achieved using same-distribution data.

We found that MIA enhances the performance of MEA on certain datasets (MRPC and RTE) when employed as a data filter. Furthermore, on most datasets, data generated by MDIA and filtered through MIA can effectively increase the success rate of MDIA. Both of these experimental outcomes are correlated with the success of MIA

H Experimental Results for LLM

Membership Inference Attack for LLM (LMIA): Table 2 shows the AUC of membership inference attacks on LLM under different threat models. From the table, we can see that SPV-MIA (based on memorization) consistently outperforms all baseline methods across all LLM with different architectures and fine-tuning datasets. The underwhelming attack performance of LiRA and Neighbour Attack reveals their inability to be effectively applied to practical LLMs. This phenomenon verifies the claim that existing MIAs designed for LMs (based on overfitting) can not handle LLMs with large-scale parameters. Upon further analysis, we can also draw the following conclusion: The privacy risk caused by MIAs on LLMs is positively correlated with the overall NLP performance of the model itself.

Model	CHIP-CTC		KUAKE-QIC	
	Original	ACC	Original	ACC
Llama2	0.793	0.782	0.800	0.788
Qwen	0.796	0.772	0.815	0.806

Table 23: Attack Results of the LMEA-I (D^{sha}, P-Tuning v2).

Model Inversion Attack for LLM (LMDIA): Table 3 shows the number of memorized examples (out of 50 candidates) that we identify using each of the three text generation strategies and six membership inference techniques. In total across all strategies, we identify 315 unique memorized training examples from among the 900 possible candidates, for an aggregate true positive rate of 35%. From the table, we can also see that the attacker's success rate using cross-domain data is higher than that of the no-data assumption. Out of privacy concerns, we do not display the recovered text in our paper, specific examples can be found on our GitHub.

Attribute Inference Attack for LLM (LAIA): Tables 8, 9 and 7 present the results of LMIA under different threat models. From the tables, we can see that the capability of attribute inference attacks is correlated with the size of the model; the larger the model, the higher the success rate of the attribute inference attacks. Age and gender are attributes that are easier to infer, whereas education and occupation are more difficult to infer, likely because the latter represent more complex attribute information.

Model Extraction Attack for LLM (LMEA): In Table 4, we present the experimental results for LMEA-G. The data from Table 4 indicate that

Model	CHIP-CTC		KUAKE-QIC	
	Original	ACC	Original	ACC
Llama2	0.793	0.788	0.800	0.801
Qwen	0.796	0.785	0.815	0.820

Table 24: Attack Results of the LMEA-I (D^{par}, P-Tuning v2).

Qwen is more vulnerable to the LMEA-G attack, consequently revealing more of its knowledge. In Tables 5, 23, 6, and 24, we display the experimental results for LMEA-I. We can observe that, regardless of the threat model and fine-tuning strategy employed, attackers can easily transfer knowledge from the target model to the extraction model. Based on the comparison of the two types of LMEA mentioned above, we can infer that general LLMs possess a stronger capacity to resist LMEA compared to domain-specific LLMs.