DPDLLM: A Black-box Framework for Detecting Pre-training Data from Large Language Models

Baohang Zhou[1], Zezhong Wang[2], Lingzhi Wang[2], Hongru Wang[2], Ying Zhang[1,*], Kehui Song[1], Xuhui Sui[1], Kam-Fai Wong[2]
[1] College of Computer Science, VCIP, TMCC, TBI Center, Nankai University, China
[2] The Chinese University of Hong Kong, China
zhoubaohang@dbis.nankai.edu.cn, {zzwang, lzwang, hrwang}@se.cuhk.edu.hk
yingzhang@nankai.edu.cn, kfwong@se.cuhk.edu.hk

Abstract

The success of large language models (LLM) benefits from large-scale model parameters and large amounts of pre-training data. However, the textual data for training LLM can not be confirmed to be legal because they are crawled from different web sites. For example, there are copyrighted articles, personal reviews and information in the pre-training data for LLM which are illegal. To address the above issue and develop legal LLM, we propose to detect the pre-training data from LLM in a pure black-box way because the existing LLM services only return the generated text. The previous most related works are the membership inference attack (MIA) on machine learning models to detect the training data from them. But the existing methods are based on analyzing the output probabilities of models which are unrealistic to LLM services. To tackle the problem, we firstly construct the benchmark datasets by collecting textual data from different domains as the seen and unseen pre-training data for LLMs. Then, we investigate a black-box framework named DPDLLM, with the only access to the generated texts from LLM for detecting textual data whether was used to train it. In the proposed framework, we exploit GPT-2 as the reference model to fit the textual data and feed the generated text from LLM into it to acquire sequence probabilities as the significant feature for detection. The experimental results on the benchmark datasets demonstrate that DPDLLM is effective on different popular LLMs and outperforms the existing methods.

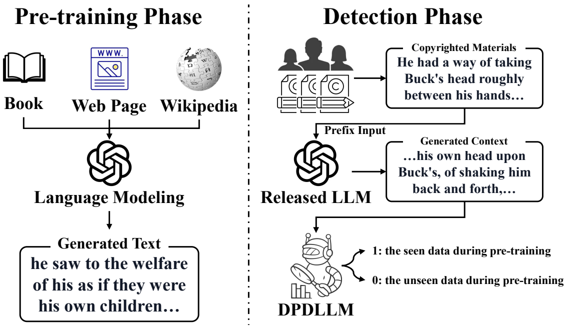
Pre-training Phase
Book
Web Page
Wikipedia
Language Modeling
Generated Text
he saw to the welfare of his as if they were his own children…
Detection Phase
Copyrighted Materials
He had a way of taking Buck's head roughly between his hands…
Prefix Input
Released LLM
Generated Context
…his own head upon Buck's, of shaking him back and forth,…
DPDLLM
1: the seen data during pre-training
0: the unseen data during pre-training

Figure 1: The schematic diagrams of pre-training phase of LLMs and the detection phase for copyrighted materials on LLMs.

1 Introduction

Language modeling (LM) is a major approach to advance natural language processing systems by modeling the likelihood of sequence of words occurring in a given context (Zhao et al., 2023). Considering to combine the neural networks with language modeling for generating human-like texts, the pre-trained language models (PLM) like: BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) were designed with different pre-training strategies. And they are effective in adapting to various downstream tasks. As the size of PLMs increases, researchers find that large-sized PLMs display the emergent abilities which are different from smaller PLMs (Wei et al., 2022). Therefore, the research community denotes the large-size PLMs as large language models (LLMs) which present the impressive performance in conversation with humans (Wang et al., 2023). The representative LLMs include the open-source model LLaMA (Touvron et al., 2023a) and the closed-source model GPT-3.5 and 4 (OpenAI, 2023). During the pre-training phase, the LLMs are trained with various corpora using the language modeling objective, as illustrated in Figure 1. However, the pre-training corpus may contain the illegal data like: copyrighted articles, personal reviews or information because they are crawled from different web sites (Touvron et al., 2023a). To develop legal LLMs, we propose to detect the unauthorized pre-training data whether is exploited to train LLMs.

The previous related works are the membership inference attack (MIA) on machine learning models to detect whether a given sample was present

[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 644–653
August 11-16, 2024 ©2024 Association for Computational Linguistics

in a target model's training data or not (Shokri et al., 2017). Mireshghallah et al. (2022) proposed the likelihood ratio attack for PLMs to measure the training data leakage. Mattern et al. (2023) firstly applied the data augmentations for the attacks against to PLMs via neighbourhood comparison. Considering to investigate the MIA on LLMs, Shi et al. (2023) constructed the benchmark WIKIMIA and designed the grey-box approach to detection based on output probabilities of LLMs. However, there are two main challenges to detect pre-training data from LLMs according to the existing approaches. Firstly, the above studies mainly exploit the output losses or probabilities of LLMs as detection features (Mattern et al., 2023; Shi et al., 2023), but it is unrealistic to obtain the output values except for generated texts from the existing LLM services. Secondly, the existing MIA approaches (Mireshghallah et al., 2022) always need to train a shadow model with the same architecture of target model, and they are time-consuming and inefficient way to detect LLMs because of the large model size and corpus.

To overcome the above disadvantages, we propose a black-box framework named **DPDLLM** [1] to detect pre-training data from LLMs. Based on the actual scenario, we require that the black-box approach only obtains the generated texts, without access to losses or probabilities of LLMs, as depicted in Figure 1. The DPDLLM framework consists of two main components: the reference model and the classifier. Given texts to be detected, we use the reference model to fit them for memorizing the texts. Then, the prefix of the detection text, such as the first half, is fed into LLMs to obtain the generated text. Ultimately, we input them into the reference model to obtain the language modeling sequence probabilities as important features. If the detection text was employed to pre-train LLMs, the corresponding generated text is similar to the raw text and the reference model will assign the high probabilities to it because of the memorization on it. The features of detection texts are fed into the classifier to predict if they were utilized during pre-training. To assess different detection methods on popular LLMs, we construct the large-scale benchmark datasets by collecting the seen and unseen texts during pre-training LLMs. The experimental results demonstrate the superiority of the DPDLLM framework on the proposed benchmark datasets. And the further analysis of different factors on the datasets verifies the robustness of the proposed method. The contributions of the manuscript can be summarized as follows:

- We analyze the disadvantages of the existing MIA-based methods on detecting pre-training data from LLMs. And to assess different detection methods on popular LLMs, we construct the large-scale benchmark datasets in different domains.

- To overcome the shortcomings, we are the first one to propose a black-box framework DPDLLM containing the reference model and the classifier. Without access of the output losses or probabilities from LLMs, the framework can detect the copyrighted materials solely based on the generated texts of LLMs.

- The experimental results and further analysis on the proposed benchmark datasets present that the DPDLLM framework outperforms existing baselines in terms of both effectiveness and robustness.

2 Related Work

2.1 Membership Inference Attacks on Language Models

Membership inference attack (MIA) was proposed to identify whether a given sample was used in training a target model or not (Shokri et al., 2017). Language models are trained with large amount of corpus where there might be personally identifiable information crawled from the public Internet. Therefore, the MIA methods could be exploited to extract the privacy information from language models (Carlini et al., 2021). One class of MIAs depends on analyzing the output values like: losses or probabilities of samples fed into target models for determining membership (Song and Raghunathan, 2020). Mattern et al. (2023) proposed a data augmentation-based attack method to compare the losses of the generated neighbour sentences and those of the original sample under the target model by computing their difference. Shi et al. (2023) selected the $k\%$ tokens with the minimum probabilities and calculated the negative log-likelihood values of them as features to detect the pre-training data for LLMs. The other class of MIAs need to train shadow models for analyzing the difference of

[1] When ready, the code will be published at `https://github.com/ZovanZhou/DPDLLM`

behaviors between target models with them. And the architecture of shadow models is always the same with that of target models. Mireshghallah et al. (2022) exploited an energy-based language model to calculate the likelihood radio between signals from both the target model and a reference model to decide the membership of a sample.

Compared with the existing studies, we focus on applying MIA on detecting pre-training data from LLMs and propose the black-box framework which do not require losses or probabilities from LLMs and time on training shadow models.

2.2 Data Contamination

Data contamination occurs when instances from the evaluation set are inadvertently included in the training dataset (Dodge et al., 2021). Recent studies have revealed that data contamination is a pervasive issue within widely-used NLP benchmark datasets (Touvron et al., 2023b; OpenAI, 2023). With the growing popularity of LLMs, it is imperative to rigorously evaluate and mitigate this problem to ensure that model performance is assessed accurately. Some existing approaches utilize sub-string matching between training samples and validation ones to identify data contamination. Brown et al. (2020) employed n-gram overlap as a method to detect contamination. Meanwhile, in the study by Chowdhery et al. (2022) on PaLM, a sample is regarded as contaminated if at least 70% of its 8-grams are present at least once in the training set. Li (2023) suggested a new approach to measure contamination levels using perplexity of LLMs, even without access to the complete training set. Other approaches try to detect contamination by extracting memorized samples with prompting LLMs with dataset names (Sainz et al., 2023) or partial contents (Golchin and Surdeanu, 2023).

While the data contamination methods can identify the evaluation samples within the training set, they are not efficient in detecting pre-training data from LLMs. This is mainly due to the impracticality of obtaining the entire corpus used for pre-training LLMs. Therefore, we propose the DPDLLM framework, which requires a small part of corpus data to train the classifier for detecting pre-training data from LLMs.

3 Methodology

Before getting into the details of the proposed framework, we formalize the problem of detecting pre-training data from LLMs. Given the target large language model f_θ pre-trained on the seen dataset D_s, the task objective is to learn a detector h for classifying the detection text $x : h(x, f_\theta) \rightarrow \{0, 1\}$ where $x \in D_s \cup D_u$ and D_u represents the unseen dataset which is not used to pre-train LLM. The training and test sets are denoted as D_{tr} and D_{te} respectively. And we formulate the training set as $D_{tr} = \{(T_i, y_i)\}_{i=1}^{|D_{tr}|}$ where T_i is the text of i-th sample and $y_i \in \{0, 1\}$ is the task label representing whether the text was utilized to pre-train LLM or not. In our setting of the task, the detector has access to the LLMs as a pure black box, and we can only acquire the generated texts but not token probabilities from LLMs. And we can obtain a small part of seen dataset because the report claimed the details of the entire pre-training corpus for popular LLMs (Touvron et al., 2023a). After training the detector on the training set, we evaluate it on the test set D_{te} and the samples of test set are disjoint with those of training set $D_{tr} \bigcap D_{te} = \varnothing$.

The DPDLLM framework for detecting pre-training data from LLMs is illustrated in Figure 2. The framework consists of the reference model and the classifier, and is trained to detect pre-training data from target LLMs. To mimic the realistic scenario, we demand that the LLMs can be accessed in a black-box way to acquire the generated contents and the reference model can be utilized to calculate the token probabilities in a white-box way. Given the texts to be detected, the reference model is trained with them for memorizing the whole texts. To obtain the memorization on detection texts of the target LLM, we feed the prefix of the texts, such as the first half, into the LLM to generate the postscripts. The trained reference model is employed on the whole generated texts to calculate the language modeling sequence probabilities as significant detection features, and we use the Gaussian Naive Bayes classifier to fit them for detection.

3.1 Memorizing Detection Texts

The pre-training stage is exploited to enhance the language modeling ability of large language models (LLM) (Zhao et al., 2023). And the popular LLMs are trained with a large amount of corpus from different domains. Considering to reduce training the shadow model same to the target LLM with the large corpus, we only need to mimic the language modeling of LLM on detection texts. We denote the detection text with N words as $T = (w_1, w_2, \ldots, w_N)$. The language modeling

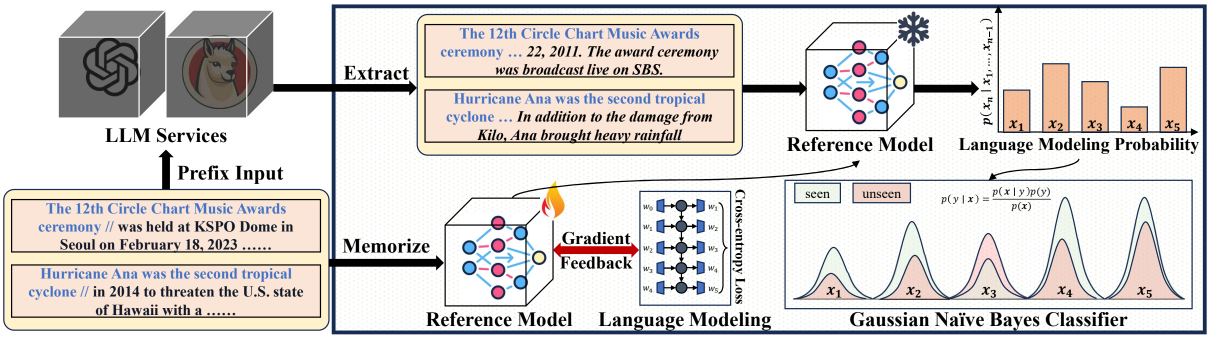
LLM Services
Prefix Input
The 12th Circle Chart Music Awards ceremony // was held at KSPO Dome in Seoul on February 18, 2023 ……
Hurricane Ana was the second tropical cyclone // in 2014 to threaten the U.S. state of Hawaii with a ……
Extract
The 12th Circle Chart Music Awards ceremony … 22, 2011. The award ceremony was broadcast live on SBS.
Hurricane Ana was the second tropical cyclone … In addition to the damage from Kilo, Ana brought heavy rainfall
Reference Model
$p(x_n \mid x_1, \dots, x_{n-1})$
x_1
x_2
x_3
x_4
x_5
Language Modeling Probability
Memorize
Reference Model
Gradient Feedback
w_0
w_1
w_2
w_3
w_4
w_5
Cross-entropy Loss
Language Modeling
seen
unseen
$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$
Gaussian Naïve Bayes Classifier

Figure 2: The DPDLLM framework for detecting pre-training data from large language models. The LLM Services can be accessed in a black-box way that users feed texts into them for generating contents. In the DPDLLM framework, the prefix texts of detection texts are input into LLMs to generate the *postscripts*.

of LLM is to generate the next words based on the previous texts. Due to the large size of LLM, the target model can memorize the pre-training data with the language modeling objective. Therefore, we employ the reference model g_ϕ to memorize the detection texts. And the language modeling objective function is defined as follows:

$$\max_{\phi} \sum_{i=1}^{N} \log p(x_i \mid x_1, x_2, \dots, x_{i-1}) \tag{1}$$

Considering that the size of the reference model is far smaller than that of the target LLM, we fine-tune the reference model by fitting it on the detection texts. Therefore, the reference model can memorize the detection texts effectively as the same as the target LLM.

3.2 Extracting Memorization of Large Language Model

During the pre-training stage, the target model f_θ fitted on the pre-training data. Therefore, the LLM can memorize the pre-training data and generate the similar texts based on the same previous contents. Given the detection text T, we want to know whether it was utilized to pre-train the target LLM. Based on the goal of pre-training stage, we can extract the memorization of LLM on the detection text. We denote the prefix text of the given detection text T as $T' = (x_1, x_2, \dots, x_{\lfloor \frac{N}{2} \rfloor})$. To extract the memorization of LLM, we feed the prefix text into the target LLM to generate postscripts. The extraction process can be simplified as follows:

$$\hat{T} = f_\theta(T') = (x_1, \dots, x_{\lfloor \frac{N}{2} \rfloor}, \hat{x}_{\lfloor \frac{N}{2} \rfloor + 1}, \dots, \hat{x}_N) \tag{2}$$

where $\hat{T}$ is the memorization text of LLM and $(\hat{x}_{\lfloor \frac{N}{2} \rfloor + 1}, \dots, \hat{x}_N)$ is the generated postscripts.

3.3 Detection Feature and Classifier

The grey-box methods utilized the losses or probabilities from target models as the detection features (Li, 2023; Shi et al., 2023). However, we can not acquire the above features because of the black-box access to the target LLM. Given the text T to be detected, we can obtain the memorization text $\hat{T}$ by Equation 2. Besides, we have the reference model g_ϕ to memorize T with the same language modeling objective function. If the target LLM was pre-trained with T, the memorization text $\hat{T}$ from LLM is similar to the raw text. And the reference model will assign the memorization text $\hat{T}$ with high probabilities because of the memorization on the detection text T. Therefore, we exploit the token probabilities from the reference model as the detection features. The language modeling sequential probabilities of the memorization text are formulated as:

$$\begin{aligned} \mathbf{O} = g_\phi(\hat{T}) = \{o_i \mid i = 1, 2, \dots, N\}, \\ o_i = p(x_i \mid x_1, x_2, \dots, x_{i-1}) \end{aligned} \tag{3}$$

where $o_i \in \mathbb{R}$ represents the conditional probability of the i-th token in $\hat{T}$. Given the training set D_{tr}, we can calculate the detection features $\{\mathbf{O}_i, y_i\}_{i=1}^{|D_{tr}|}$ with the above steps.

Considering that the token probabilities are independent and identical distribution, we assume that the distributions of detection features follow the normal distribution. Therefore, we propose to make use of Gaussian Naive Bayes (GNB) classifier to map the features to the detection labels by modeling the distributions of features. To classify the sample with N-dimension feature into category set c, the parameters of GNB are defined as: $\tau = \{(\mu_k, \sigma_k) \mid k \in c\}$. For the data of the detection task, we can estimate the parameters of the

Algorithm 1 DPDLLM

```
Input: Training set $D_{tr}$, Test set $D_{te}$, LLM $f_\theta$
       Reference model $g_\phi$, GNB classifier $h_\tau$
Output: Results on test set $Y = \{\tilde{y}_i\}_{i=1}^{|D_{te}|}$.
/* Training Procedure */
Update $\phi_{tr} \leftarrow \phi$ with $D_{tr}$ by Equation 1
for $i = 1, 2, \dots, |D_{tr}|$ do
    Obtain $T_i^{tr}$ as the $i$-th sample in $D_{tr}$
    Extract $\hat{T}_i \leftarrow f_\theta(T_i^{tr})$ by Equation 2
    Calculate $\mathbf{O}_i \leftarrow g_{\phi_{tr}}(\hat{T}_i)$ by Equation 3
end for
Update $\tau$ with $\{\mathbf{O}_i, y_i\}_{i=1}^{|D_{tr}|}$ by Equation 4
/* Evaluation Procedure */
Update $\phi_{te} \leftarrow \phi$ with $D_{te}$ by Equation 1
for $i = 1, 2, \dots, |D_{te}|$ do
    Obtain $T_i^{te}$ as the $i$-th sample in $D_{te}$
    Extract $\hat{T}_i \leftarrow f_\theta(T_i^{te})$ by Equation 2
    Calculate $\mathbf{O}_i \leftarrow g_{\phi_{te}}(\hat{T}_i)$ by Equation 3
    Calculate $\tilde{y}_i$ with $\mathbf{O}_i$ by Equation 5
end for
return $Y = \{\tilde{y}_i\}_{i=1}^{|D_{te}|}$
```

classifier as follows:

$$\mu_k = \frac{1}{|D_{tr}|} \sum_{i=1, y_i=k}^{|D_{tr}|} \mathbf{O}_i,$$
$$\sigma_k = \sqrt{\frac{1}{|D_{tr}|} \sum_{i=1, y_i=k}^{|D_{tr}|} (\mathbf{O}_i - \mu_k)^2} \tag{4}$$

where $\{\mu_k, \sigma_k\} \in \mathbb{R}^N$ and $c = \{0, 1\}$ represents the two categories. To inference the detected data, we maximize the posterior probability by the Bayes theorem as follows:

$$\begin{aligned} \tilde{y} &= \arg\max_c \log(p(y = c \mid \mathbf{O})) \\ &= \arg\max_c \log(p(y = c) \prod_{i=1}^{N} p(o_i \mid y = c)). \end{aligned} \tag{5}$$

The parameters of the classifier are fixed after training and used to test on the evaluation data.

3.4 Training and Evaluation Procedures

In the DPDLLM framework, the reference model is required to memorize the detection texts and output the probabilities of the corresponding memorization texts extracted from the target LLM f_θ. And the classifier is fed with the detection features of texts and identifies their labels for detecting them whether were utilized to pre-train the target model.

Therefore, given the training set D_{tr}, we firstly train the reference model $g_{\phi_{tr}}$ and the classifier h_τ with the optimized parameters τ during the training procedure. To evaluate the framework on the test set D_{te}, we consider that the detection texts from two sets are dis-joint and should also train the reference model $g_{\phi_{te}}$ to memorize the texts in D_{te}. During the evaluation procedure, we obtain the detection features based on the reference model $g_{\phi_{te}}$ and use the classifier h_τ to recognize the type of the texts to be detected. The overall algorithm of DPDLLM is illustrated in Algorithm 1.

Dataset	WikiMIA2	BookMIA	Wiki-SPGC	WikiMIA2-SPGC
Length	64	512	64	64
# Train set	2,252	4,935	1,928	3,214
# Test set	2,252	4,935	1,929	3,215
Distribution (S%:U%)	51% : 49%	50% : 50%	50% : 50%	66% : 34%
Domain				
Book		S, U	S	S
Website	S, U		U	S, U

Table 1: The statistical information of the four benchmark datasets. The "Domain" item shows the data sources of the corresponding dataset. "S" means the seen data source that were used to pre-train the target LLMs and "U" means the unseen one that were not employed during pre-training.

4 Experimental Setup

To detect the pre-training data from LLMs, we investigate the open-source LLMs including: Pythia-2.8B (Biderman et al., 2023), LLaMA-7B and LLaMA-13B (Touvron et al., 2023a) as target models. Based on the reports of the above LLMs' training details, we can construct the detection datasets.

Datasets. To evaluate the different detection methods, Shi et al. (2023) proposed the benchmark datasets WikiMIA and BookMIA. However, the above datasets are limited to the small size and domains. Therefore, we propose the new benchmark datasets with large size and more domains. The detailed information of the benchmark datasets is shown in Table 1.

WikiMIA2: Considering that there are only 542 samples in WikiMIA, we extend it as WikiMIA2 by collecting newest contents from websites including: Wikipedia and news. Based on the Wikipedia page [2] of events occurred in 2023, we crawled the texts and the reference websites of news in it as the unseen data. And we follow the report of

[2] https://en.wikipedia.org/wiki/2023

WikiMIA2												
Methods	Pythia-2.8B				LLaMA-7B				LLaMA-13B			
	Prec.	Reca.	F1	AUC	Prec.	Reca.	F1	AUC	Prec.	Reca.	F1	AUC
PPL	51.5	97.9	37.0	50.2	51.5	97.9	37.0	50.2	51.3	96.4	37.3	49.8
Zlib	51.2	**98.2**	35.5	49.6	51.2	**98.2**	35.5	49.6	51.4	**98.4**	36.5	50.1
Lowercase	51.5	96.2	38.3	50.2	51.5	96.2	38.3	50.2	50.9	94.0	37.7	49.2
Average PROB	54.7	73.0	53.4	54.5	55.1	76.2	53.9	55.3	54.3	80.6	51.9	54.5
MIN-K% PROB	54.3	81.5	51.6	54.5	58.6	57.2	57.3	57.3	56.6	65.9	56.0	56.2
DPDLLM	**66.7**	75.7	**67.9**	**67.9**	**71.8**	73.2	**71.4**	**71.4**	**73.0**	68.5	**70.8**	**70.8**

BookMIA												
Methods	Pythia-2.8B				LLaMA-7B				LLaMA-13B			
	Prec.	Reca.	F1	AUC	Prec.	Reca.	F1	AUC	Prec.	Reca.	F1	AUC
PPL	82.1	3.2	37.2	51.3	**88.6**	3.2	37.2	51.4	**94.8**	4.5	38.6	52.1
Zlib	77.6	2.4	36.3	50.9	86.8	2.7	36.7	51.2	92.9	3.2	37.3	51.5
Lowercase	61.6	22.6	49.6	54.4	61.3	19.5	47.9	53.7	63.4	17.4	47.1	53.8
Average PROB	78.0	**78.5**	78.4	78.4	77.6	**78.1**	78.0	78.0	80.3	**79.1**	80.1	80.1
MIN-K% PROB	74.3	54.3	67.5	68.0	83.7	50.0	69.1	70.2	84.5	53.4	71.1	71.9
DPDLLM	**82.7**	77.6	**80.9**	**80.9**	81.9	77.7	**80.5**	**80.5**	83.9	77.9	**81.7**	**81.6**

Table 2: Performance comparison of different detection methods on benchmark datasets including: WikiMIA2 and BookMIA. We investigate the large language models including: Pythia-2.8B, LLaMA-7B and LLaMA-13B as the target models to detect the pre-training data from them.

LLMs (Touvron et al., 2023a) to use the previously dumped pages of Wikipedia as the seen data which were employed to pre-train them.

BookMIA: Shi et al. (2023) constructed the dataset by extracting 100 random 512-word snippets from 100 books of the Books3 corpus (Gao et al., 2021) as the seen data and collecting 50 new books with first editions in 2023 as the unseen data.

Wiki-SPGC: To construct the multi-domain dataset, we select books of Gutenberg corpus (Gerlach and Font-Clos, 2020) as the seen data and take the unseen data of WikiMIA2 into account.

WikiMIA2-SPGC: To mimic the seen data from multi-domains, we merge WikiMIA2 with the books of Gutenberg corpus as the whole dataset. And it is a hard benchmark to evaluate the detection methods under the multi-domain scenario.

Baselines. We take existing detection methods as our baselines and utilize them to extract detection features for classification. The widely used feature is the perplexity (PPL) (Li, 2023) of the sample fed into language models (LM). The other PPL-based methods include comparing the sample PPL to zlib compression entropy (Zlib) or to the lowercased sample PPL (Lowercase) (Carlini et al., 2021). Besides, we take the sentence probability from LM as detection feature and apply the average operation on it (Average PROB). Shi et al. (2023) proposed to select the $k\%$ tokens with the minimum probability and compute the average log-likelihood of them as detection feature.

In the experiments, we select GPT-2 as the reference model and train it for 15 epochs to memorize detection texts. For different baselines, we extract their specific features to optimize classifiers based on training set and evaluate them on test set.

5 Experimental Results

5.1 Main Results

To compare the performance of different detection methods on WikiMIA2 and BookMIA, we investigate the large language models including: Pyhia-2.8B, LLaMA-7B and LLaMA-13B as target models. As illustrated in Table 2, the proposed DPDLLM framework can always gain the best results including: F1 and AUC scores on the two datasets and three LLMs. The PPL-based detection methods including: PPL, Zlib and Lowercase are not useful on the two benchmark datasets. The memorization texts extracted from LLMs differ from the original detection texts. And the PPL reflects the loss of the reference model that is trained using the detection texts. Consequently, the PPL values of the memorization texts fed into the refer-

Methods	WikiMIA2				Wiki-SPGC				WikiMIA2-SPGC			
	Prec.	Reca.	F1	AUC	Prec.	Reca.	F1	AUC	Prec.	Reca.	F1	AUC
PPL	51.5	99.6	35.9	50.3	97.3	11.4	44.9	55.6	65.9	**100.0**	52.4	50.0
Zlib	51.4	**99.7**	35.3	50.1	**100.0**	3.32	37.0	51.7	65.9	**100.0**	52.4	50.0
Lowercase	54.2	80.5	51.6	54.3	94.9	**99.2**	**96.9**	**96.9**	66.3	94.9	55.4	50.8
Average PROB	55.1	76.2	53.9	55.3	89.7	81.7	86.2	86.2	65.9	**100.0**	52.4	50.0
MIN-K% PROB	58.6	57.2	57.3	57.3	87.6	66.3	78.2	78.5	65.9	**100.0**	52.4	50.0
DPDLLM	**71.8**	73.2	**71.4**	**71.4**	90.7	85.0	88.2	88.2	**73.7**	78.0	**66.6**	**62.1**

Table 3: Performance comparison of different detection methods on multi-domain scenarios. We investigate the large language model LLaMA-7B as the target models to detect the pre-training data from it.

Methods	WikiMIA2		BookMIA	
	Pythia-2.8B	LLaMA-13B	Pythia-2.8B	LLaMA-13B
DPDLLM	67.9	70.8	80.9	81.6
Reference Model				
w/o fine-tuning	66.8	67.2	73.6	75.1
Classifier				
LR	65.2	70.0	76.8	77.7
MLP	64.4	65.0	73.3	74.3

Table 4: The ablation study of DPDLLM on the benchmark datasets. The results of each method are AUC scores. "LR" represents the logistic regression model and "MLP" represents the multi-layer perceptron model.

ence model are not discriminative features for detecting pre-training data. Besides, the probability-based methods including: Average PROB and MIN-K% PROB achieve better performances than the PPL-based methods. If the original texts used to pre-train target models are memorized by the reference model, the memorization texts extracted from LLMs will be assigned high token probabilities. Therefore, the token probabilities from the reference model are the efficient features to detect pre-training data. Average PROB and MIN-K% PROB gain worse results than the proposed method. Because the average operation compresses the whole sequence probability into one value, and the significant features are reduced by it. And DPDLLM takes the sequence probability into account and models the distributions of it for detection.

5.2 Analysis on Multi-Domains

Considering that the detection texts may come from different domains, we construct the multi-domain benchmark datasets including: Wiki-SPGC and WikiMIA2-SPGC. As illustrated in Table 3, the overall metrics like: F1 and AUC scores of Wiki-SPGC are higher than those of the other datasets. Because Wiki-SPGC consists of the unseen texts from websites and the seen ones from books, and there is significant difference between the two kinds of texts for the easy detection. And Lowercase method can gain the best results on Wiki-SPGC because the significant difference between website and book texts lies in the writing convention of upper and lower case. Besides, WikiMIA2-SPGC is the hardest benchmark dataset becuase it contains the unseen texts from websites and the seen texts from both websites and books. As shown in Table 3, the baselines are not effective on WikiMIA2-SPGC and identify the all detection texts as the seen ones which were utilized to pre-train LLMs. Because the seen texts are from different domains, and the detection features of baseline methods are not useful to train the classifier. The proposed DPDLLM framework can also gain the best results and the phenomenon verifies that the sequence probability is useful in detecting pre-training data of multi-domains from LLMs.

5.3 Ablation Study

The DPDLLM framework consists of the reference model and the classifier. To investigate the effectiveness of components in DPDLLM, we conduct the ablation study of it on the benchmark datasets and LLMs. The reference model is trained to memorize the detection texts by language modeling objective function. We utilize the reference model without fine-tuning to extract detection features for identifying pre-training data. As presented in Table 4, the results of the reference model without fine-tuning are worse than those of DPDLLM. Because the original reference model can not memorize the detection texts without training and it can not assign the reasonable probabilities to the memorization texts extracted from LLMs. Besides, we replace the Gaussian Naive Bayes (GNB) classifier with other ones. As shown in Table 4, the logistic regression (LR) model can achieve the competitive results on WikiMIA2 compared with GNB, but it gains worse results than GNB on BookMIA. This

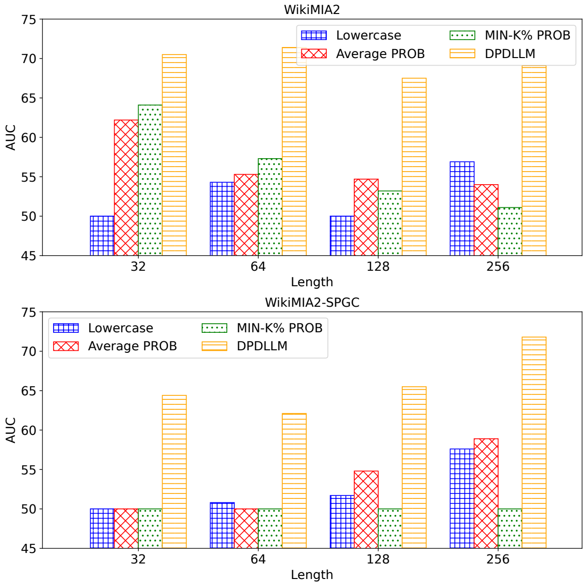

Figure 3: The performance comparison of detection methods on benchmark datasets with various lengths. The target large language model is LLaMA-7B.

phenomenon demonstrates that the text lengths of detection texts influence the results of different classifiers, and the GNB can keep the superiority on long texts compared with LR. And the multi-layer perceptron (MLP) model gains the worst results because it can not model the distributions of detection features and generalize to the test set effectively.

5.4 Influence of Text Length

To analyze the influence of text lengths to the performances of different detection methods, we construct the benchmark datasets with various lengths versions. There are four versions of WikiMIA2 and WikiMIA2-SPGC with different token lengths including: 32, 64, 128 and 256. As illustrated in Figure 3, DPDLLM framework can always achieve the best results on the benchmark datasets with different lengths. This phenomenon verifies that the sequence probability is the detection features robust to text lengths. Besides, Lowercase method can gain better results with the increase of text lengths. And the performances of Average PROB and MIN-K% PROB on WikiMIA2 decrease with the increase of text lengths. Because WikiMIA2 contains texts from the same domain and the probability-based methods struggle to use the increased texts for detection. The MIN-K% PROB method on WikiMIA2-SPGC is unable to detect multi-domain pre-training data with different lengths. This is because it only utilizes a small portion of token probabilities as detection features. In contrast, other methods that rely on sequence probabilities can achieve better results as the length of the text increases. This observation supports the idea that longer detection texts improve results of detecting multi-domain pre-training data from LLMs.

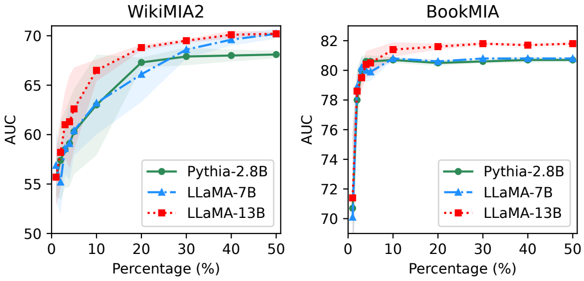

Figure 4: The performance comparison of DPDLLM on benchmark datasets including: WikiMIA2 and BookMIA with different percentages of original training data.

5.5 Influence of Training Set Size

To investigate the impact of training set size, we randomly select a portion of the original training set to train DPDLLM. We then evaluate its performance on the test set. The results, shown in Figure 4, demonstrate that DPDLLM achieves better performance as the training data size increases. Specifically, when training DPDLLM on WikiMIA2, utilizing 50% of the original training data set yields the best results. On the other hand, using only 10% of the original training data set yields the best results on BookMIA. Therefore, it is not necessary to have large amounts of annotated training data to detect pre-training data from LLMs.

6 Conclusion

In this paper, we discuss the issue of using potentially illegal pre-training data for large language models (LLMs) and analyze the disadvantages of existing detection methods. Considering to avoid training shadow models and accessing inner of LLMs, we propose the black-box framework named DPDLLM to detect pre-training data from LLMs. The proposed framework comprises a reference model and a classifier. The reference model is trained to memorize detection texts, while the classifier is optimized using detection features of memorization texts extracted from target LLMs, based on the reference model. To evaluate detection methods, we construct the benchmark datasets from different domains and the experimental results demonstrate the superiority and robustness of the proposed methods over baselines.

7 Limitations

The experimental results demonstrate that the existing detection methods are not effective on multi-domain dataset WikiMIA2-SPGC. Therefore, we should investigate more significant detection features and efficient detection models to identify multi-domain pre-training data from LLMs. Besides, in the proposed framework, we need to make use of a small part of seen data which were utilized to pre-train LLMs. And it is more convenient to combine the detection methods with the zero-shot learning to reduce utilizing the distribution of pre-training data for LLMs.

Acknowledgements

We thank the anonymous reviewers for the valuable comments on our manuscript. This research is supported by the National Natural Science Foundation of China (No. 62272250, 62302243, U22B2048, 62077031), and the Natural Science Foundation of Tianjin, China (No. 22JCJQJC00150, 23JCQNJC01960).

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. 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 2021, August 11-13, 2021*, pages 2633–2650. USENIX Association.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. *CoRR*, abs/2308.08493.

Yucheng Li. 2023. Estimating contamination via perplexity: Quantifying memorisation in language model evaluation. *CoRR*, abs/2309.10677.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and

Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada. Association for Computational Linguistics.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

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.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023. Did chatgpt cheat on your test?

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, pages 377–390. ACM.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064, Singapore. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.