Training Data Extraction From Pre-trained Language Models: A Survey

Shotaro Ishihara
Nikkei Inc.
1-3-7, Otemachi, Chiyoda-ku, Tokyo
shotaro.ishihara@nex.nikkei.com

Abstract

As the deployment of pre-trained language models (PLMs) expands, pressing security concerns have arisen regarding the potential for malicious extraction of training data, posing a threat to data privacy. This study is the first to provide a comprehensive survey of training data extraction from PLMs. Our review covers more than 100 key papers in fields such as natural language processing and security. First, preliminary knowledge is recapped and a taxonomy of various definitions of memorization is presented. The approaches for attack and defense are then systemized. Furthermore, the empirical findings of several quantitative studies are highlighted. Finally, future research directions based on this review are suggested.

1 Introduction

Pre-trained language models (PLMs) are widely used in natural language processing. Statistical models that assign probabilities to token sequences have been studied, and large neural networks are increasingly being used for pre-training with large datasets. This scaling has led to fluent natural language generation and success in many other downstream tasks (Devlin et al., 2019). In some cases, parameter updates are not required for downstream tasks (Radford et al., 2019; Brown et al., 2020).

With increasing applications of PLMs, security concerns have increased considerably (Bender et al., 2021; Bommasani et al., 2021; Weidinger et al., 2022). Studies have revealed the risk of language models exhibiting unintentional *memorization* of training data, and occasionally outputting memorized information (Carlini et al., 2019, 2021, 2023b; Lee et al., 2023). In particular, Carlini et al. (2021) identified that personal information can be extracted by generating numerous sentences from PLMs and performing *membership inference* (Shokri et al., 2017). These attacks on PLMs are referred to as *training data extraction* and are undesirable because of privacy, decreased utility, and reduced fairness concerns (Carlini et al., 2023b). However, with the evolution of PLMs, limited progress has been achieved in addressing these concerns, and security technology is yet to mature.

This study is the first to provide a comprehensive survey of training data extraction from PLMs. Starting with the pioneering work, we reviewed more than 100 previous and subsequent studies. Specifically, we screened papers citing Carlini et al. (2021)[1] based on the relationships, the number of citations, and their acceptance. First, Section 2 presents preliminary knowledge. We then discuss several topics with the following contributions:

- A taxonomy of various **definitions of memorization** (Section 3) was presented. Training data extraction has become close to the famous security attack known as model inversion (Fredrikson et al., 2015).
- We systematize the approaches to **attack** (Section 4) and **defense** (Section 5). Furthermore, we highlight **empirical findings** (Section 6) from several quantitative evaluation studies.
- Based on the review, we suggest **future research directions** (Section 7).

2 Preliminaries about PLMs

This section describes the basics of modern PLMs. First, we explain the methodology used for training language models and generating texts. Next, the standard practical schema is introduced.

2.1 Language Models

Language models represent a probability distribution over the sequences of tokens. Based on the pre-training method, language modeling can be categorized into two types (Yang et al., 2023): *autoregressive language modeling*, which predicts words

[1] https://scholar.google.com/scholar?cites=12274731957504198296

sequentially from left to right (Bengio et al., 2000; Mikolov et al., 2010), and *masked language modeling*, which hides some parts of a sentence and fills in the gaps (Devlin et al., 2019). The former is sometimes called *causal language modeling* (Tirumala et al., 2022).

This study is focused on autoregressive language models with transformer (Vaswani et al., 2017), following many recent studies on training data extraction. Note that some studies have focused on masked language models such as BERT (Lehman et al., 2021; Mireshghallah et al., 2022a; He et al., 2022) and T5 (Carlini et al., 2023b). Most studies address pre-training rather than fine-tuning (Mireshghallah et al., 2022b).

Autoregressive language models take a series of tokens as input and output a probability distribution for the next token. We show a schema of training and generation by following Carlini et al. (2021).

Training. The following statistical model was assumed for distribution:

$$\mathbf{Pr}(x_1, x_2, \ldots, x_n),$$

where $x_1, x_2, \ldots, x_n$ is a sequence of tokens from a vocabulary using the chain rule of probability:

$$\mathbf{Pr}(x_1, x_2, \ldots, x_n) = \Pi_{i=1}^{n} \mathbf{Pr}(x_i \mid x_1, \ldots, x_{i-1}).$$

Let $f_\theta(x_i \mid x_1, \ldots, x_{i-1})$ denote the likelihood of token x_i when evaluating neural network f with parameters θ. Language models are trained to optimize the probability of the data in a training set. Formally, training involves minimizing the loss function as follows:

$$\mathcal{L}(\theta) = -\log \Pi_{i=1}^{n} f_\theta(x_i \mid x_1, \ldots, x_{i-1})$$

for each data in the training set. This setting can be qualitatively regarded as memorizing the flow of sentences in each training data.

Generating. New tokens can be generated by iterating the following process:

1. Choose $\hat{x}_{i+1} \sim f_\theta(x_{i+1} | x_1, \ldots, x_i)$.
2. Feed $\hat{x}_{i+1}$ back into the model to choose $\hat{x}_{i+2} \sim f_\theta(x_{i+2} | x_1, \ldots, \hat{x}_{i+1})$.

This decoding process continues until conditions are satisfied. The simplest is greedy decoding, selecting the most probable tokens one by one. However, studies have revealed that simply maximizing the output probability generates text that is not natural to humans (Li et al., 2016; Holtzman et al., 2020). Therefore, several approaches have been proposed for sampling from a probability distribution such as top-k sampling (Fan et al., 2018) and top-p sampling (Appendix A).

2.2 Pre-training and Fine-tuning

Prior to BERT (Devlin et al., 2019), specific models were trained for individual tasks. By contrast, in the PLMs approach, large neural networks with large datasets are pre-trained and fine-tuned for several downstream tasks. Radford et al. (2018) revealed that autoregressive language modeling is effective for PLMs with transformers. This extension, GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), can be applied to various tasks without fine-tuning by providing a few examples (in-context learning). The scaling of large models with large datasets has attracted considerable research attention (Appendix B).

PLMs exhibit a significant advantage in using datasets that match a specific domain. These models can exhibit superior performance in domain-specific tasks than larger models pre-trained on general datasets. Studies, such as BioMegatron (Shin et al., 2020), BioGPT (Luo et al., 2022), Galactica (Taylor et al., 2022), and BloombergGPT (Wu et al., 2023), have been conducted. However, the potential risk of training data extraction, especially when using sensitive datasets in pre-training, should be considered (Nakamura et al., 2020; Lehman et al., 2021; Jagannatha et al., 2021; Singhal et al., 2022; Yang et al., 2022). There are also ethical topics such as the human rights in the texts (Li et al., 2018; Ginart et al., 2019; Garg et al., 2020; Henderson et al., 2022) and plagiarism regarding copyright (Lee et al., 2023). Examples include PLMs created from contracts (Chalkidis et al., 2020; Zheng et al., 2021), clinical information (Kawazoe et al., 2021), music (Agostinelli et al., 2023), and source code (Chen et al., 2021).

3 Definitions of Memorization

Memorization is the concept that PLMs store and output information about the training data. There is a wide variety of research on memorization, with diverse definitions and assumptions. We illustrate a taxonomy of definitions in Figure 1.

3.1 Eidetic memorization

A mainstream method is *eidetic memorization* (Carlini et al., 2021) and its variations (Thomas McCoy et al., 2021; Carlini et al., 2023b; Kandpal

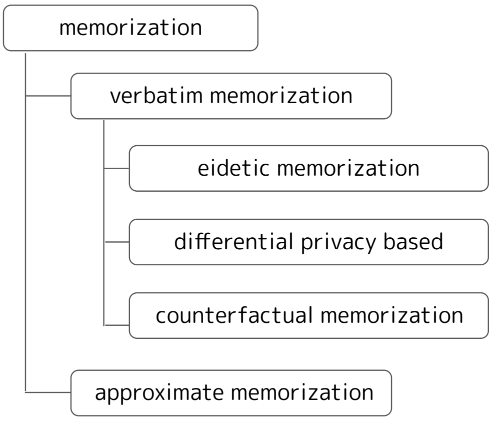

Figure 1: Taxonomy of definitions of memorization.

et al., 2022; Tirumala et al., 2022). These definitions assume that PLMs output memorized data when appropriate prompts are provided. Carlini et al. (2021) defined eidetic memorization as Definition 3.1, and in a subsequent study (Carlini et al., 2023b), they adopted the definition in Definition 3.2. They stated that eidetic memorization can be used in cases in which no prompt, whereas the subsequent definition is suitable for conditions with prompts. Some studies have adopted definitions similar to those in Definition 3.2. Examples include Tirumala et al. (2022) with a per-token definition of *exact memorization*, and Kandpal et al. (2022) with a document-level definition of *perfect memorization*.

Definition 3.1 (eidetic memorization)**.** A string s is k-eidetic memorized by PLM f_θ if a prompt p exists such that $f(p) = s$ and s appears at most k times in the training set.

Definition 3.2 (a variation of eidetic memorization)**.** A string s is k-memorized with k tokens of context from a PLM f_θ if a (length-k) string p exists such that the concatenation $[p||s]$ is contained in the training set, and f_θ produces s when prompted with p by using greedy decoding.

3.2 Differential privacy

Differential privacy (Dwork et al., 2006) is widely used in memorization, and definitions based on differential privacy have been devised (Jagielski et al., 2020; Nasr et al., 2021). Differential privacy was formulated based on the premise that removing any data from the training set should not considerably change trained models. Although this method protects the personal information of a single user, Brown et al. (2022) reported that the method cannot capture the complexity of social and linguistic data. Differential privacy is introduced as a defense approach in Section 5.2.

3.3 Counterfactual memorization

Studies have defined *counterfactual memorization* as the difference between a training data's expected loss under a model that has and has not been trained on that data (Feldman and Zhang, 2020; van den Burg and Williams, 2021). Zhang et al. (2021c) investigated this form of memorization in PLMs based on the taxonomy of human memorization in psychology.

The definition of counterfactual memorization has received limited attention in training data extraction. Carlini et al. (2023b) noted that this definition requires training thousands of models to measure privacy. Thus, evaluating PLMs becomes difficult because of their inference costs. Furthermore, Kandpal et al. (2022) remarked that this definition is not considered a privacy attack scenario because access to the training corpus is assumed. This phenomenon is related to the adversarial knowledge presented in Section 4.2.

3.4 Approximate memorization

Although the definitions of memorization thus far assume exact string matches, definitions have been proposed to relax this condition. Here, Ippolito et al. (2022) refer to definitions based on exact string matches as *verbatim memorization*. They revealed that verbatim memorization can be handled by simply adjusting the decoding method and proposed alternative definitions called *approximate memorization* that consider string fuzziness, as presented in Definition 3.3. Some methods have been proposed to calculate similarity. Ippolito et al. (2022) set the condition that $\mathrm{BLEU}(\mathrm{s}, \mathrm{g})$ (Papineni et al., 2002) is greater than 0.75. The threshold value of 0.75 was selected by qualitatively inspecting examples. Lee et al. (2022) defined that the token is memorized if it is part of a substring of 50 tokens of a string in the training data.

Definition 3.3 (approximate memorization)**.** A string s is k-approximately memorized by PLM f_θ if a (length-k) string p exists such that (s, g) satisfies certain conditions of similarity, and f_θ produces g when prompted with p.

3.5 Revisiting model inversion

Reconstructing training data from a model presents a well-known security concern called model inver-

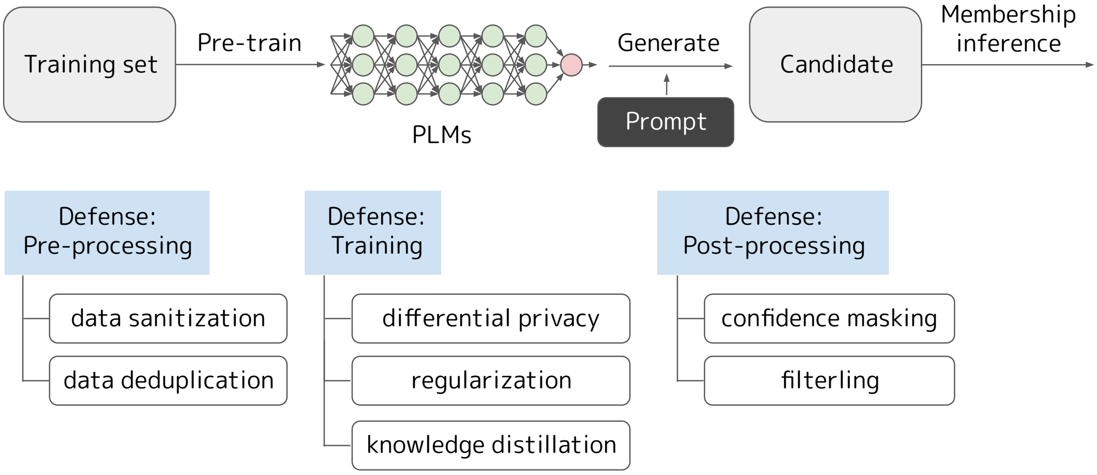

Figure 2: The procedure of training data extraction attacks and possible defenses.

sion attacks (Fredrikson et al., 2015). Carlini et al. (2021) explained that the main difference is that training data extraction does not allow fuzziness. However, this difference has decreased since the introduction of relaxed definitions of memorization. Kandpal et al. (2022) mentioned several previous studies (Carlini et al., 2019, 2021; Inan et al., 2021) as model inversion.

4 Training Data Extraction Attacks

This section systematizes the attack procedure. Most studies follow Carlini et al. (2021). They revealed that hundreds of verbatim text sequences can be extracted from the training data. Given a PLM, the procedure consists of two steps, candidate generation, and membership inference, as displayed in Figure 2.

4.1 Candidate generation

The first step is to generate numerous texts from a given PLM. Texts can be generated from PLMs using several decoding methods, as discussed in Appendix A. Here, Carlini et al. (2023b) reported that the choice of the decoding strategy does not considerably affect their experimental results. In contrast, Lee et al. (2023) observed that top-k and top-p sampling tended to extract more training data.

Another perspective is the procedure for providing prompts. Prompts are provided according to two options, giving only a special token[2] (sometimes called *no prompt*) or specific strings as prompts. Studies have constructed prompts by extracting data from the dataset considered to be used in creating PLMs. Carlini et al. (2021) randomly sampled between 5 and 10 tokens from scraped data. Carlini et al. (2023b) extracted a subset of the Pile dataset (Gao et al., 2020) in prompting GPT-Neo model family (Black et al., 2022).

4.2 Membership inference

Membership inference aims to predict whether any particular example is used to train a machine learning model (Shokri et al., 2017; Song and Shmatikov, 2019; Hisamoto et al., 2020). This result can lead directly to privacy violations. We describe membership inference on PLMs from the following five perspectives in a survey paper (Hu et al., 2022): target model, adversarial knowledge, approach, algorithm, and domain.

Target model. This study focuses on autoregressive language models as discussed in Section 2.1. Attacks on other models such as word embeddings (Song and Raghunathan, 2020; Mahloujifar et al., 2021; Meehan et al., 2022), natural language understanding (Parikh et al., 2022), text classification (Nasr et al., 2019; Zhang et al., 2022; Elmahdy et al., 2022), and image diffusion models (Carlini et al., 2023a) exist but are not covered.

Adversarial knowledge. The second perspective is the knowledge that can be handled explicitly by attackers. We describe two aspects of adversarial knowledge, namely models and training sets. The patterns of adversarial knowledge in this study are summarized in Appendix C.

Hu et al. (2022) presented the adversarial knowledge of models. The models are classified into two categories, namely white-box and black-box,

[2] Carlini et al. (2021) used `<|endoftext|>`, as indicated at `https://github.com/ftramer/LM_Memorization`.

according to accessibility (Nasr et al., 2019). Under the white-box setting, an attacker can obtain all information and use it for the attack. This includes the training procedure and the architecture and trained parameters of the target model. However, in the black-box setting, an attacker can only have limited access to the target model. Hu et al. (2022) classified the black-box setting into three parts, namely full confidence scores, top-k confidence scores, and prediction labels only. They differ in the extent of access an attacker has to the PLMs output. The setting of full confidence scores assumes a situation in which the training process of the model is unknown, but all outputs for any given input are available. Therefore, an attacker can obtain prediction labels with probabilities and calculate the loss. The setting of top-k confidence scores indicates that an attacker can obtain several candidates of the output. The scope of the attack is restricted because losses cannot be calculated. Another setting provides only labels without prediction values (Choquette-Choo et al., 2021; Zhu et al., 2023). Many web services with PLMs, such as DeepL[3] and ChatGPT[4], only allow users to view labels for the model output.

Furthermore, we describe the adversarial knowledge of the training sets. In the white-box setting, the training set is stated and publicly available. The most harmful attacks are black box setups that do not assume access to the training set. Such attacks include PLMs created by private datasets. In some cases, the data are partially publicly available. Such cases include the ones wherein only the beginning of the news article is available for free, certain editions are accessible, and some articles have been made private over time. Although the data itself are not partially published, substrings can be inferred in the hidden private data using a priori knowledge (Henderson et al., 2018; Carlini et al., 2019). Examples are prompts like *"Bob's phone number is"* and *"Alice's password is"*.

We must be aware of scenarios in which the dataset and PLMs are unwillingly leaked and become public. Adversarial knowledge is immediately converted to the white-box level. For example, even if a web service with PLMs trained on a private dataset provides users with only a string, it is crucial to discuss risks when both the dataset and the PLMs are unintentionally made public.

Approach. Hu et al. (2022) divided the membership inference approaches into three categories, namely classifier-based (Shokri et al., 2017; Song and Shmatikov, 2019), metric-based (Bentley et al., 2020; Choquette-Choo et al., 2021; Song and Mittal, 2021), and differential comparisons (Hui et al., 2021). For example, in shadow training (Shokri et al., 2017; Song and Shmatikov, 2019), a primary classifier-based method, additional training is assumed in the model (white-box settings). Some metric-based methods can be applied to realistic black-box settings.

In studies of training data extraction from PLMs, *perplexity* is often used for metrics of membership inference (Carlini et al., 2019, 2021). Given a sequence of tokens $x_1, \ldots, x_n$, the perplexity is defined as:

$$\mathcal{P} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} \log f_\theta(x_i | x_1, \ldots, x_{i-1})\right)$$

Algorithm. The fourth perspective is whether the algorithm is centralized or federated. Federated learning approaches have received considerable attention in privacy protection research (Melis et al., 2019; Nasr et al., 2019; Lee et al., 2021; Kairouz et al., 2021). However, focusing on training data extraction, the mainstream approach is based on centralized methods as of April 2023.

Domain. Text datasets are rooted in various domains, as described in Section 2.2. Clinics are a crucial research field that involves handling of highly confidential information. Lehman et al. (2021) recovered patient names and their associated conditions from PLMs using electronic clinical records. Jagannatha et al. (2021) demonstrated that patients with rare disease profiles may be highly vulnerable to higher privacy leakages through experiments using PLMs of clinical data. Many other domains require careful processing, such as contracts (Yin and Habernal, 2022) and source code[5]. A discussion of the right to be forgotten in the legal and news industries has emerged (Li et al., 2018; Ginart et al., 2019; Garg et al., 2020; Henderson et al., 2022). Therefore, it should be ensured that PLMs do not unintentionally become digital archives.

Publicly available datasets do not necessarily indicate that they are completely independent of the risk of training data extraction from PLMs. The context in which the information is shared

[3] https://www.deepl.com/translator

[4] https://openai.com/blog/chatgpt/

[5] https://github.blog/2021-06-30-github-copilot-research-recitation/

should be known to respect privacy (Dourish, 2004; Nissenbaum, 2009). Nissenbaum's contextual integrity (Nissenbaum, 2009) states that a change in any one of five characteristics (data subject, sender, recipient, information type, and transmission principle) may alter privacy expectations. Brown et al. (2022) emphasized the importance of PLMs only with data explicitly intended for public use. The Italian Data Protection Authority issued a statement[6] on March 2023 in accordance with the European General Data Protection Regulation (GDPR) against OpenAI, the provider of ChatGPT, for their data processing.

5 Training Data Extraction Defenses

This section systematizes approaches to defense. We can mitigate privacy risks before, during, and after creating PLMs as displayed in Figure 2. The classification was reconstructed using references (Hu et al., 2022; Huang et al., 2022; Jagielski et al., 2023). Extensive studies have been conducted on the hazardous generation of PLMs (Kurita et al., 2020; Mei et al., 2022; Levy et al., 2022; Ouyang et al., 2022; Carlini et al., 2023c). However, this study focused on training data extraction.

5.1 Pre-processing

First, pre-processing the training set is considered.

Data sanitization. The simplest solution is to identify and remove any text that conveys personal information (Ren et al., 2016; Continella et al., 2017; Vakili et al., 2022). However, as noted in Section 4.2, privacy depends on the context, and determining privacy from the string alone is difficult. Brown et al. (2022) proposed that data sanitization is only useful for removing context-independent, well-defined, static pieces of personal information from the training set.

Data deduplication. Studies have indicated that data deduplication mitigates the memorization of PLMs (Allamanis, 2019; Kandpal et al., 2022; Lee et al., 2022). This method is more efficient than methods that train models and is expected to be a practical solution. Empirical findings on data deduplication are presented in Section 6.2.

5.2 Training

The second method is a pre-training strategy.

Differential privacy. Applying differential privacy (Dwork et al., 2006) methods for providing data privacy guarantees in machine learning models has attracted considerable research attention. Differential privacy is a data protection measure that is designed to ensure that providing data does not reveal much information about the user. However, applying these algorithms (e.g., DP-SGD (Abadi et al., 2016) and DP-FedAvg (Ramaswamy et al., 2020)) to PLMs is challenging. Performance degradation and increased computation and memory usage are the primary concerns.

To address this problem, a framework has been proposed for training models in two steps (Yu et al., 2021, 2022; Li et al., 2022; He et al., 2023)[7]. In the framework, large amounts of non-private datasets are used for pre-training to obtain general features; next, additional training is applied with a sensitive dataset using a differential privacy algorithm. Downey et al. (2022) reported that the differential privacy approach is effective in preventing memorization, despite its computational and model performance costs. Note that Tramèr et al. (2022) summarized a critical view. They argued that publicly accessible datasets are not free from privacy risks because they contain information that is unintentionally released to the public. Therefore, discussing whether private information that we want to hide is contained in the public dataset is essential. It is known that understanding the semantic guarantee of differential privacy is difficult when private data is involved (Cummings et al., 2021).

Another barrier to applying differential privacy to PLMs is the requirement of defining secret boundaries even though text data are not binary. Studies have considered various levels of granularity, from individual tokens or words to sentences, documents, or even the entire user dataset (McMahan et al., 2018; Levy et al., 2021; Lukas et al., 2023).

Regularization. Regularization is a well-known approach for suppressing overfitting in machine learning models. The memorization of models is typically associated with overfitting (Yeom et al., 2018; Zhang et al., 2021b). Therefore, regularization during training that reduces overfitting can be used as a measure of membership inference (Hu et al., 2022). Mireshghallah et al. (2021) proposed a regularization method regarding the memoriza-

[6] https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870847

[7] A study has also appeared that applies these algorithms to in-context learning settings (Panda et al., 2023).

tion of PLMs and claimed usefulness compared with differential privacy methods. Some studies have constrained the representation of neural networks by the information bottleneck layer (Alemi et al., 2017; Henderson and Fehr, 2023).

Pre-training large neural networks has distinctive tendencies compared with common machine learning. A single data in the training set is not used for too many epochs in pre-training and is sometimes used for less than one epoch. Furthermore, Carlini et al. (2021) reported that a characteristic of PLM memorization is the emergence of training data with an abnormally lower loss than the average. Tirumala et al. (2022) revealed that large language models can memorize most of their data before overfitting and tend not to forget much information through the training process. Biderman et al. (2023) have focused on the training process and attempted to predict the memorization of PLMs.

Knowledge distillation. Another approach is knowledge distillation (Hinton et al., 2015), in which the output of a large teacher model is used to train a small student model. Shejwalkar and Houmansadr (2021) revealed that knowledge distillation can be used to restrict an attacker's direct access to a private training set, which considerably reduces membership information leakage.

5.3 Post-processing

The third step is to post-process the PLMs output.

Confidence masking. Limiting the output of PLMs is a simple but effective defense mechanism. For example, confidence masking can be used for adjusting adversarial knowledge, as presented in Section 4.2 and Appendix C.

Filtering. Filtering the output of PLMs before providing them to users is crucial. Identifying items to be filtered incurs a cost, and ensuring diversity remains challenging. Perez et al. (2022) proposed a method to automatically identify test cases by extracting potentially dangerous outputs by detailing prompts using various PLMs.

6 Empirical Findings

This section presents empirical findings on training data extraction from PLMs. Initial studies were limited to qualitative evaluations, but subsequent studies (Lee et al., 2022; Kandpal et al., 2022; Ippolito et al., 2022; Tirumala et al., 2022; Downey et al., 2022; Carlini et al., 2023b; Lee et al., 2023) have focused on quantitative evaluations.

In particular, based on one of the first comprehensive quantitative studies (Carlini et al., 2023b), we report on the impact of the model size, the string duplication in the training set, and the length of prompts. They used various sizes of GPT-Neo model family (Black et al., 2022), which are the autoregressive language models pre-trained by the Pile dataset (Gao et al., 2020). Four model sizes, namely 125 million, 1.3 billion (B), 2.7 B, and 6 B parameters, were considered. The number of duplicate strings was determined by analyzing the Pile dataset. A subset of 50,000 sentences from the Pile dataset was used for evaluation, and the distribution of duplicates was considered. The beginning of each sentence was cut out at a certain number of tokens and considered as a prompt. The amount of memorization was calculated as the fraction of generations that exactly reproduce the true string for their prompt averaged over all prompts and sequence lengths.

6.1 Larger models memorize more

Carlini et al. (2023b) revealed that a near-perfect log-linear relationship exists such that the larger the model size is, the more strings are memorized. Numerically, a ten-fold increase in the model size increased the amount of memorization by 19 ppt. For comparison, they performed the same analysis with the GPT-2 model family. The amount of memorization was 40 % for 1.3 B GPT-neo compared with 6 % for the GPT-2 of the same size. This phenomenon implied the effect of memorization of the training data, not just the model size.

Carlini et al. (2023b) used the definition of verbatim memorization, and Ippolito et al. (2022) confirmed similar results with the definition of approximate memorization. Although not sufficiently quantitative, initial studies (Carlini et al., 2019; Zhang et al., 2021b) have provided preliminary evidence. Tirumala et al. (2022) and Lee et al. (2023) also revealed that larger models memorize more.

6.2 Duplicate strings are memorized

Carlini et al. (2023b) reported that a clear log-linear trend exists between the number of duplicates and the amount of memorization. They measured the amount of memorization for each bucket with duplicate counts ranging from 2 to 900. Kandpal et al. (2022) and Lee et al. (2022) also revealed that duplication in the training set of PLMs relates to the likelihood of memorizing strings and proposed that deduplication mitigates training data extraction. However, memorization can occur even with only a few duplicates, and deduplication cannot prevent it completely. Chang et al. (2023) reported that the degree of memorization of ChatGPT and GPT-4 (OpenAI, 2023) was related to the frequency of the passages that appeared on the web.

6.3 Longer prompts extract more

Carlini et al. (2023b) revealed that the amount of memorization increases with the length of the prompt. For example, the amount of memorization by the 6 B model was 33 % for 50 tokens, compared with 65 % for 450 tokens. This experiment was inspired by the findings of Carlini et al. (2019). They suggested that setting the maximum prompt length available to users considerably reduces the risk of training data extraction.

7 Conclusion & Future Directions

We have reviewed over 100 papers for the first comprehensive survey on training data extraction from PLMs. The final section provides suggestions for future research directions. We hope that this study highlights the importance of training data extraction from PLMs and accelerates the discussion.

7.1 Is memorization always evil?

Most studies did not distinguish the degree of danger of memorized strings (Lee et al., 2020). Ideally, the undesirable memorization of telephone numbers and email addresses must be separated from the acceptable memorization. Huang et al. (2022) was among the first to differentiate between memorization and association in PLMs. They concluded that the risk of specific personal information being leaked is low because PLMs cannot semantically associate personal information with their owners.

The boundary between memorization and knowledge of PLMs remains ambiguous with the definition of approximate memorization (Ippolito et al., 2022; Lee et al., 2022). Deduplication of training sets, which is considered useful in Sections 5 and 6, leads to the elimination of helpful knowledge. Therefore, we must consider what memorization is (Haviv et al., 2022) and balance the security concerns with the model performance, depending on the final application. The definition of counterfactual memorization introduced in Section 3.3 incorporated psychological findings that could be useful despite its challenges.

7.2 Toward broader research fields

Discussing the handling of the fuzziness of a string is important. Ippolito et al. (2022) stated that the current definition of approximate memorization focuses on English, and different considerations are required for other conditions such as non-English languages. In addition, they suggested two research areas that could help improve the definition: image generation memorization and plagiarism detection. Images are more difficult to generate than text for matching exactly with the original. Therefore, fuzzy memorization has been investigated and measured. Fredrikson et al. (2015), which proposed the model inversion attack, used face recognition in images as the subject of their experiments. Studies have used metrics that consider image similarity (Zhang et al., 2020; Haim et al., 2022; Balle et al., 2022). Furthermore, the trend toward pre-training in both images and language (Lu et al., 2019; Li et al., 2020) should be considered. The limitations of the definition of verbatim textual matching have been discussed in plagiarism detection research (Roy et al., 2009; Potthast et al., 2010). Similarities are explored from multiple perspectives, including word changes, shuffling, and paraphrasing.

7.3 Evaluation schema

Room for ingenuity exists in the construction of evaluation sets. Establishing a schema for quantitative evaluation, which has received considerable attention, is critical. Studies mentioned in Sections 4 and 6 have created evaluation sets by extracting a subset of the training set. Sampling is essential because of inference time limitations. However, we must be careful to see if there are other factors to consider besides the distribution of the number of duplicates to avoid bias due to sampling.

Evaluation metrics for the training data extraction are open for discussion. Carlini et al. (2022) postulated that the ideal evaluation metric must be based on realistic attack scenarios, whereas most studies on membership inference measure the average accuracy rate. They proposed that membership inference should be evaluated by the true positive rate with a low false positive rate. The Training Data Extraction Challenge[8] measures attack speed as well as recall and precision.

[8] https://github.com/google-research/lm-extraction-benchmark

Limitations

First, this study focused on PLMs in training data extraction, particularly autoregressive language models. Other target models, such as masked language models (described in Section 2.1) and word embeddings (noted in Section 4.2), require another discussion. Additionally, due to prioritization constraints, the discussion on other topics, including model inversion attacks and the federated learning approach, was limited. However, these areas are established and can be supplemented by other studies (Fredrikson et al., 2015; Zhang et al., 2021a).

Second, in practical applications of PLM, it is necessary to audit not only security but also various other aspects such as performance degradation (Mökander et al., 2023). There are a number of security concerns beyond training data extraction (noted in Section 5). There are also papers discussing performance degradation of PLMs over time (Ishihara et al., 2022).

Finally, this comprehensive survey is based on information as of April 2023. Studies on training data extraction from PLMs have primarily focused on natural language processing and security. These domains are undergoing rapid changes. Therefore, some of the content may become obsolete in the near future.

Ethics Statement

The privacy concerns regarding training data extraction from PLMs were reviewed to help mature discussions in academia and industry. Of course, its purpose is not to promote these attacks.

Studies on PLMs tend to focus on the English language, which is the language used by the majority of people in the world, and the same is true for training data extraction. Therefore, this study focused on English. As indicated in Section 7.2, research on other languages is encouraged.

Acknowledgements

We would like to thank Editage (www.editage.com) for English language editing.

References

Martin Abadi, Andy Chu, Ian Goodfellow, et al. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA. Association for Computing Machinery.

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, et al. 2023. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, et al. 2017. Deep variational information bottleneck. In *Proceedings of the 5th International Conference on Learning Representations*.

Miltiadis Allamanis. 2019. The adverse effects of code duplication in machine learning models of code. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, Onward! 2019, pages 143–153, New York, NY, USA. Association for Computing Machinery.

Balle, Cherubin, and Hayes. 2022. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, volume 0, pages 1138–1156.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, et al. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Jason W Bentley, Daniel Gibney, Gary Hoppenworth, et al. 2020. Quantifying membership inference vulnerability via generalization gap and other model metrics. *arXiv preprint arXiv:2009.05669*.

Stella Biderman, Usvsn Sai Prashanth, Lintang Sutawika, et al. 2023. Emergent and predictable memorization in large language models. *arXiv preprint arXiv:2304.11158*.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, et al. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2280–2292, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Steve Chien, Milad Nasr, et al. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914.

Nicholas Carlini, Jamie Hayes, Milad Nasr, et al. 2023a. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, et al. 2023b. Quantifying memorization across neural language models. In *Proceedings of the 11th International Conference on Learning Representations*.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, et al. 2023c. Poisoning Web-Scale training datasets is practical. *arXiv preprint arXiv:2302.10149*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, et al. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramèr, Eric Wallace, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Kent K Chang, Mackenzie Cramer, Sandeep Soni, et al. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. *arXiv preprint arXiv:2305.00118*.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, et al. 2021. Label-only membership inference attacks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1964–1974. PMLR.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Andrea Continella, Yanick Fratantonio, Martina Lindorfer, et al. 2017. Obfuscation-resilient privacy leak detection for mobile apps through differential analysis. In *Proceedings 2017 Network and Distributed System Security Symposium*, Reston, VA. Internet Society.

Rachel Cummings, Gabriel Kaptchuk, and Elissa M Redmiles. 2021. "I need a better description": An investigation into user expectations for differential privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pages 3037–3052, New York, NY, USA. Association for Computing Machinery.

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.

Paul Dourish. 2004. What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1):19–30.

C M Downey, Wei Dai, Huseyin A Inan, et al. 2022. Planting and mitigating memorized content in Predictive-Text language models. *arXiv preprint arXiv:2212.08619*.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, et al. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg.

Adel Elmahdy, Huseyin A. Inan, and Robert Sim. 2022. Privacy leakage in text classification a data extraction approach. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 13–20, Seattle, United States. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: discovering the long tail via influence estimation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, number Article 242 in NIPS'20, pages 2881–2891, Red Hook, NY, USA. Curran Associates Inc.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 2015 ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 1322–1333, New York, NY, USA. Association for Computing Machinery.

Leo Gao, Stella Biderman, Sid Black, et al. 2020. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Sanjam Garg, Shafi Goldwasser, and Prashant Nalini Vasudevan. 2020. Formalizing data deletion in the context of the right to be forgotten. In *Advances in Cryptology – EUROCRYPT 2020*, pages 373–402. Springer International Publishing.

Antonio A Ginart, Melody Y Guan, Gregory Valiant, et al. 2019. Making AI forget you: data deletion in machine learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, NIPS'19, pages 3518–3531, Red Hook, NY, USA. Curran Associates Inc.

Niv Haim, Gal Vardi, Gilad Yehudai, et al. 2022. Reconstructing training data from trained neural networks. In *Advances in Neural Information Processing Systems*.

Adi Haviv, Ido Cohen, Jacob Gidron, et al. 2022. Understanding transformer memorization recall through idioms. *arXiv preprint arXiv:2210.03588*.

Jiyan He, Xuechen Li, Da Yu, et al. 2023. Exploring the limits of differentially private deep learning with group-wise clipping. In *Proceedings of the 11th International Conference on Learning Representations*.

Xuanli He, Lingjuan Lyu, Chen Chen, and Qiongkai Xu. 2022. Extracted BERT model leaks more information than you think! In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1530–1537, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Henderson and Fabio James Fehr. 2023. A VAE for transformers with nonparametric variational information bottleneck. In *Proceedings of the 11th International Conference on Learning Representations*.

Peter Henderson, Mark S Krass, Lucia Zheng, et al. 2022. Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset. In *36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, et al. 2018. Ethical challenges in Data-Driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 123–129, New York, NY, USA. Association for Computing Machinery.

Tom Henighan, Jared Kaplan, Mor Katz, et al. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Sorami Hisamoto, Matt Post, and Kevin Duh. 2020. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Transactions of the Association for Computational Linguistics*, 8:49–63.

Ari Holtzman, Jan Buys, Li Du, et al. 2020. The curious case of neural text degeneration. In *Proceedings of the 8th International Conference on Learning Representations*.

Hongsheng Hu, Zoran Salcic, Lichao Sun, et al. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys*, 54(11s).

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bo Hui, Yuchen Yang, Haolin Yuan, et al. 2021. Practical blind membership inference attack via differential comparisons. In *28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021*. The Internet Society.

Huseyin A Inan, Osman Ramadan, Lukas Wutschitz, et al. 2021. Training data leakage analysis in language models. In *3rd Privacy-Preserving Machine Learning Workshop*.

Daphne Ippolito, Florian Tramèr, Milad Nasr, et al. 2022. Preventing verbatim memorization in language models gives a false sense of privacy. *arXiv preprint arXiv:2210.17546*.

Shotaro Ishihara, Hiromu Takahashi, and Hono Shirai. 2022. Semantic shift stability: Efficient way to detect performance degradation of word embeddings and pre-trained language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 205–216, Online only. Association for Computational Linguistics.

Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Membership inference attack susceptibility of clinical language models. *arXiv preprint arXiv:2104.08305*.

Matthew Jagielski, Om Thakkar, Florian Tramèr, et al. 2023. Measuring forgetting of memorized training examples. In *Proceedings of the 11th International Conference on Learning Representations*.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing differentially private machine learning: how private is private SGD? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, number Article 1862 in NIPS'20, pages 22205–22216, Red Hook, NY, USA. Curran Associates Inc.

Peter Kairouz, H Brendan McMahan, Brendan Avent, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, et al. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, et al. 2021. A clinical specific BERT developed using a huge japanese clinical text corpus. *PloS one*, 16(11):e0259763.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806, Online. Association for Computational Linguistics.

Hongkyu Lee, Jeehyeong Kim, Seyoung Ahn, et al. 2021. Digestive neural networks: A novel defense strategy against inference attacks in federated learning. *Computers and Security*, 109(C).

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jooyoung Lee, Thai Le, Jinghui Chen, et al. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3637–3647, New York, NY, USA. Association for Computing Machinery.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959, Online. Association for Computational Linguistics.

Daniel Levy, Ziteng Sun, Kareem Amin, et al. 2021. Learning with user-level privacy. In *Advances in Neural Information Processing Systems*.

Sharon Levy, Emily Allaway, Melanie Subbiah, Lydia Chilton, Desmond Patton, Kathleen McKeown, and William Yang Wang. 2022. SafeText: A benchmark for exploring physical safety in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2407–2421, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11336–11344.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Tiffany Li, Eduard Fosch Villaronga, and Peter Kieseberg. 2018. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, 34(2):304.

Xuechen Li, Florian Tramer, Percy Liang, et al. 2022. Large language models can be strong differentially private learners. In *Proceedings of the 10th International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, et al. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nils Lukas, Ahmed Salem, Robert Sim, et al. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*.

Renqian Luo, Liai Sun, Yingce Xia, et al. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Saeed Mahloujifar, Huseyin A Inan, Melissa Chase, et al. 2021. Membership inference on word embedding and beyond. *arXiv preprint arXiv:2106.11384*.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, et al. 2018. Learning differentially private recurrent language models. In *Proceedings of the 6th International Conference on Learning Representations*.

Casey Meehan, Khalil Mrini, and Kamalika Chaudhuri. 2022. Sentence-level privacy for document embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3367–3380, Dublin, Ireland. Association for Computational Linguistics.

Alex Mei, Anisha Kabir, Sharon Levy, Melanie Subbiah, Emily Allaway, John Judge, Desmond Patton, Bruce Bimber, Kathleen McKeown, and William Yang Wang. 2022. Mitigating covertly unsafe text within natural language systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2914–2926, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, et al. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706.

Tomas Mikolov, Martin Karafiát, Lukas Burget, et al. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. fit.vutbr.cz.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022a. 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.

Fatemehsadat Mireshghallah, Huseyin Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. 2021. Privacy regularization: Joint privacy-utility optimization in LanguageModels. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3799–3807, Online. Association for Computational Linguistics.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022b. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *arXiv preprint arXiv:2302.08500*.

Yuta Nakamura, Shouhei Hanaoka, Yukihiro Nomura, et al. 2020. KART: Parameterization of privacy leakage scenarios from pre-trained language models. *arXiv preprint arXiv:2101.00036*.

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)*, pages 739–753.

Milad Nasr, Shuang Song, Abhradeep Thakurta, et al. 2021. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, volume 0, pages 866–882.

Helen Nissenbaum. 2009. *Privacy in Context*. Stanford University Press.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Ashwinee Panda, Tong Wu, Jiachen T Wang, et al. 2023. Differentially private In-Context learning. *arXiv preprint arXiv:2305.01639*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rahil Parikh, Christophe Dupuy, and Rahul Gupta. 2022. Canary extraction in natural language understanding models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 552–560, Dublin, Ireland. Association for Computational Linguistics.

Ethan Perez, Saffron Huang, Francis Song, et al. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Coling 2010: Posters*, pages 997–1005, Beijing, China. Coling 2010 Organizing Committee.

Alec Radford, Karthik Narasimhan, Tim Salimans, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Swaroop Ramaswamy, Om Thakkar, Rajiv Mathews, et al. 2020. Training production language models without memorizing user data. *arXiv preprint arXiv:2009.10031*.

Jingjing Ren, Ashwin Rao, Martina Lindorfer, et al. 2016. ReCon: Revealing and controlling PII leaks in mobile network traffic. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '16, page 361–374. Association for Computing Machinery.

Chanchal K Roy, James R Cordy, and Rainer Koschke. 2009. Comparison and evaluation of code clone detection techniques and tools: A qualitative approach. *Science of Computer Programming*, 74(7):470–495.

Virat Shejwalkar and Amir Houmansadr. 2021. Membership privacy for machine learning models through knowledge transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9549–9557.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, Online. Association for Computational Linguistics.

Reza Shokri, Marco Stronati, Congzheng Song, et al. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Shaden Smith, Mostofa Patwary, Brandon Norick, et al. 2022. Using DeepSpeed and megatron to train Megatron-Turing NLG 530b, a Large-Scale generative language model. *arXiv preprint arXiv:2201.11990*.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, pages 377–390, New York, NY, USA. Association for Computing Machinery.

Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in Text-Generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 196–206, New York, NY, USA. Association for Computing Machinery.

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

Yixuan Su, Tian Lan, Yan Wang, et al. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, et al. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

R Thomas McCoy, Paul Smolensky, Tal Linzen, et al. 2021. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *arXiv preprint arXiv:2111.09509*.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, et al. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. 2022. Considerations for differentially private learning with Large-Scale public pretraining. *arXiv preprint arXiv:2212.06470*.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Gerrit van den Burg and Chris Williams. 2021. On memorization in probabilistic deep generative models. *Advances in Neural Information Processing Systems*, 34:27916–27928.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, et al. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 214–229, New York, NY, USA. Association for Computing Machinery.

Shijie Wu, Ozan Irsoy, Steven Lu, et al. 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, et al. 2023. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *arXiv preprint arXiv:2304.13712*.

Xi Yang, Aokun Chen, Nima PourNejatian, et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, et al. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282.

Ying Yin and Ivan Habernal. 2022. Privacy-preserving models for legal natural language processing. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 172–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Da Yu, Saurabh Naik, Arturs Backurs, et al. 2022. Differentially private fine-tuning of language models. In *Proceedings of the 10th International Conference on Learning Representations*.

Da Yu, Huishuai Zhang, Wei Chen, et al. 2021. Large scale private learning via low-rank reparametrization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12208–12218. PMLR.

Chen Zhang, Yu Xie, Hang Bai, et al. 2021a. A survey on federated learning. *Knowledge-Based Systems*, 216:106775.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, et al. 2021b. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, et al. 2021c. Counterfactual memorization in neural language models. *arXiv preprint arXiv:2112.12938*.

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

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, et al. 2020. The secret revealer: Generative model-inversion attacks against deep neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Lucia Zheng, Neel Guha, Brandon R Anderson, et al. 2021. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, ICAIL '21, pages 159–168, New York, NY, USA. Association for Computing Machinery.

Tianqing Zhu, Dayong Ye, Shuai Zhou, et al. 2023. Label-only model inversion attacks: Attack with the least information. *IEEE Transactions on Information Forensics and Security*, 18:991–1005.

A Type of Decoding

Two classes of methods, namely deterministic and stochastic, are used for decoding (Su et al., 2022). In the deterministic method, the most probable tokens based on the probability distribution of the model are used. Greedy methods and beam searches are widely used. However, studies have revealed that simply maximizing the output probability generates text that is not natural to humans (Li et al., 2016; Holtzman et al., 2020). Therefore, several approaches have been proposed for sampling from a probability distribution. Stochastic methods include top-k sampling (Fan et al., 2018), top-p sampling, and nucleus sampling (Holtzman et al., 2020), in which samples are extracted from the lexical subset. A method to adjust the probability distribution using the temperature parameter was used to increase the diversity of the generated texts (Ackley et al., 1985).

In the candidate generation step in Section 4.1, texts can be generated from PLMs using several decoding methods. Some studies adopted a greedy method (Carlini et al., 2023b). Others used top-k sampling (Carlini et al., 2021; Lee et al., 2022) and tuned the temperature (Carlini et al., 2021) to increase the diversity of the generated texts.

B Scaling Law for Language Models

Building PLMs requires large datasets. Studies have proposed models with larger parameters pre-trained with large datasets (Smith et al., 2022; Chowdhery et al., 2022). Experimental results revealed the existence of a scaling law (Kaplan et al., 2020; Henighan et al., 2020). This study suggested that the performance of language models using the transformer improves as the model size, dataset size, and amount of computation increase. Villalobos et al. (2022) cautioned that the data available for pre-training language models may be exhausted in the near future.

C Patterns of Adversarial Knowledge

Table 1 presents the patterns of adversarial knowledge of the models and Table 2 details the patterns of adversarial knowledge of the training set. These tables provide specific patterns. For example, white-box for models indicates PLMs published on platforms such as Hugging Face[9] with training explanations, which can be downloaded. As discussed in Section 4.2, two main types, namely white and black boxes, exist. In black-box settings, several patterns depend on the situation. Table 1 reveals the classification of the black-box proposed by Hu et al. (2022): full confidence scores, top-k confidence scores, and prediction labels. In Table 2,

[9] https://huggingface.co/models

Adversarial knowledge	Model or the output	Pattern
white-box	all	Models are available with proper explanations.
black-box	full confidence scores	All outputs of models are available.
	top-k confidence scores	Top-k outputs of models are available.
	prediction label only	Only prediction labels are available.

Table 1: Adversarial knowledge of models and patterns.

Adversarial knowledge	Training set	Pattern
white-box	all	Dataset used for training is stated and publicly available.
black-box	partial	Dataset used for training is stated but not available.
		Dataset used for training is stated and partially available.
	nothing	Dataset used for training is not stated.

Table 2: Adversarial knowledge of training sets and patterns.

several possible patterns of adversarial knowledge are presented on training sets.