Understanding PII Leakage in Large Language Models: A Systematic Survey

Shuai Cheng[1,5] , **Zhao Li**[1,2] , **Shu Meng**[1,5] , **Mengxia Ren**[1] , **Haitao Xu**[1,5*] ,
Shuai Hao[3] , **Chuan Yue**[4] , **Fan Zhang**[1]
[1]Zhejiang University
[2]Hangzhou Yugu Technology
[3]Old Dominion University, Norfolk, VA, U.S.A.
[4]Colorado School of Mines, Golden, CO, U.S.A.
[5]The State Key Laboratory of Blockchain and Data Security, Zhejiang University
{cs36, mengshu, haitaoxu, fanzhang}@zju.edu.cn, lzjoey@gmail.com
shao@odu.edu, {mengxiaren,chuanyue}@mines.edu

Abstract

Large Language Models (LLMs) have demonstrated exceptional success across a variety of tasks, particularly in natural language processing, leading to their growing integration into numerous facets of daily life. However, this widespread deployment has raised substantial privacy concerns, especially regarding personally identifiable information (PII), which can be directly associated with specific individuals. The leakage of such information presents significant real-world privacy threats. In this paper, we conduct a systematic investigation into existing research on PII leakage in LLMs, encompassing commonly utilized PII datasets, evaluation metrics, and current studies on both PII leakage attacks and defensive strategies. Finally, we identify unresolved challenges in the current research landscape and suggest future research directions.

1 Introduction

In recent years, large language models (LLMs) have achieved notable advancements, enabling their broad integration into diverse real-world applications. However, this rapid adoption has also amplified security risks, particularly concerning privacy leakage. LLMs are primarily trained on extensive, publicly available datasets from the Internet, including personal blogs, online forums, Wikipedia, and institutional websites [Radford *et al.*, 2019; Ouyang *et al.*, 2022], which often contain substantial amounts of unauthorized personal information. Due to the absence of robust privacy protection measures in most datasets, there exists a significant risk of exposing personal privacy information, the scale of which can be considerable. This threat becomes particularly alarming when LLMs are capable of accurately generating specific personal private information, commonly referred to as personally identifiable information (PII) [McCallister *et al.*, 2010].

PII encompasses personal data such as names, email addresses, phone numbers, occupations, home addresses, educational backgrounds, and even social security numbers or private passwords [Lukas *et al.*, 2023]. The inadvertent leakage of PII embedded in the training data of LLMs can lead to severe consequences, including identity theft, fraud, and cyberattacks. Furthermore, PII included in user prompts may be memorized and unintentionally disclosed by LLMs [Staab *et al.*, 2024]. More critically, adversaries can actively exploit query-based attacks by crafting prompts to elicit PII outputs from the LLM, thereby enabling PII extraction. As a result, protecting LLMs and users from PII leakage has emerged as a critical challenge in both research and practical deployment.

While substantial research has been conducted on privacy leakage in LLMs, investigations specifically targeting PII leakage remain in their early stages. Two categories of privacy leakage research have involved the exposure of PII. The first category primarily explores the utilization of jailbreak techniques to circumvent the security constraints and ethical guidelines of language models, thereby obtaining unauthorized information [Deng *et al.*, 2024; Shen *et al.*, 2024]. This line of research focuses on developing methods to prompt LLMs to respond to PII-related inquiries without refusal [Zou *et al.*, 2023; Yu *et al.*, 2024], rather than generating verifiably authentic PII. The second category concentrates on the leakage of training data within LLMs, which may include PII. However, this body of work is constrained by the fact that only a relatively small fraction of the extracted PII is genuine, as the extraction of PII demands a significantly higher level of precision compared to the extraction of general training data.

Figure 1 provides an overview of the growing number of studies on PII leakage from 2021 to 2024, encompassing both attack and defense perspectives. Specifically, the figure enumerates the models involved and the corresponding timelines of each study. It shows that studies on PII leakage have covered both open-source models (e.g., GPT-2 and BERT) and closed-source models (e.g., GPT-3.5 and GPT-4). However, a comprehensive investigation into PII leakage in LLMs has yet to be conducted. Similarly, existing surveys on privacy leakage [Yan *et al.*, 2024; Yao *et al.*, 2024] have not provided a thorough analysis of PII leakage, making it challenging to gain a holistic understanding of the overall landscape of this field. To address this gap, this paper aims to deliver an exten-

[*] Haitao Xu is the corresponding author.

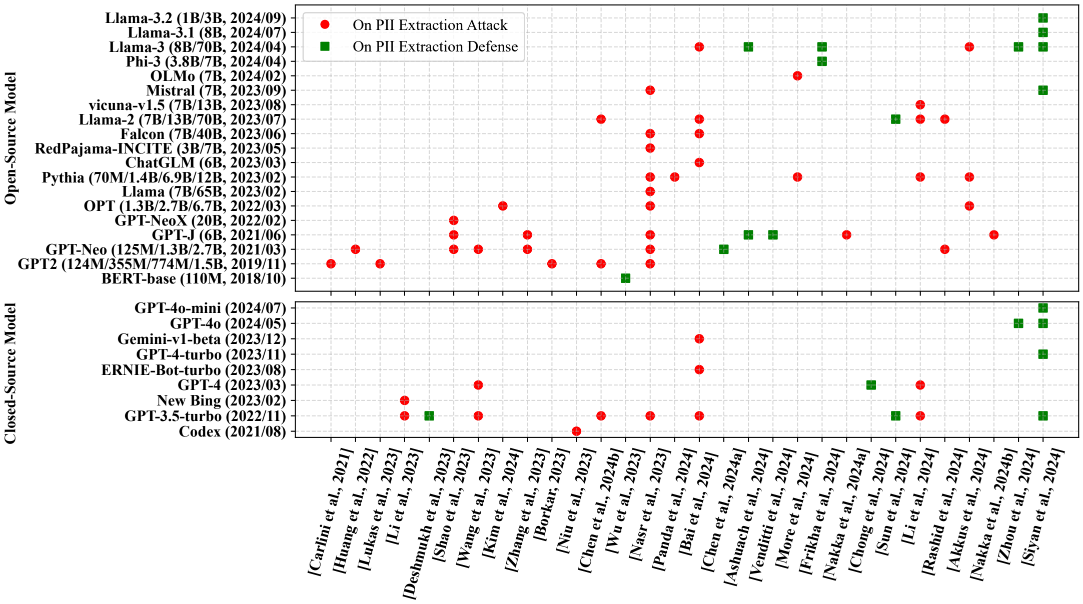

Figure 1: Overview of PII Leakage Studies. The X-axis shows relevant studies from 2021 to 2024. The Y-axis lists the models evaluated, divided into open-source (upper section) and closed-source (lower section). Open-source models are labeled with parameter sizes and release dates, ordered by release time. Closed-source models are marked only by release dates, as their parameter sizes are not publicly available.

sive survey of PII leakage issues in LLMs, with a particular focus on attack and defense methods for PII extraction. Our primary contributions are as follows:

- We introduce the commonly used datasets and evaluation metrics for PII leakage and analyze their application scenarios in PII leakage research.
- We conduct a systematic and comprehensive survey of PII leakage attack and defense methods, presenting core techniques and target LLMs of each study. We categorize these methods, analyze their strengths and limitations, and summarize their evaluation results for comparative analysis.
- We discuss the limitations of current PII leakage research and propose several future research directions and perspectives, providing insights and inspiration for further studies.

The remainder of this paper is structured as follows: §2 introduces the background of PII leakage; §3 outlines the commonly used datasets and evaluation metrics for PII leakage; §4 reviews existing techniques for PII leakage attacks; §5 presents defense methods against PII leakage; §6 explores future research directions; and §7 concludes the paper.

2 Background

In this section, we provide definitions and classifications of PII and PII leakage in LLMs and introduce two model capabilities of LLMs that can lead to PII leakage.

2.1 PII

Personally identifiable information (PII) refers to any data that can potentially identify a specific individual [McCallister *et al.*, 2010]. PII can be categorized into two distinct types: *Direct Identifiers* and *Quasi-Identifiers*. Direct identifiers are data elements that, on their own, can directly identify a specific individual, such as email addresses, phone numbers, and social security numbers. In contrast, quasi-identifiers are data points that, while not individually sufficient for identification, can be combined with other information to potentially identify an individual, such as name, dates of birth, or zip code.

2.2 PII Leakage

PII leakage occurs when an individual's sensitive personal information is disclosed without their consent. In the context of LLMs, PII leakage can occur through the following scenarios:

- **Training Data Leakage via Adversarial Prompts.** An attacker crafts carefully designed prompts to induce the LLM to reproduce large portions of its training data verbatim, from which PII can be extracted [Carlini *et al.*, 2021; Carlini *et al.*, 2023].
- **PII Extraction via Targeted and Non-Targeted Queries.** An attacker can design queries to prompt the LLM to disclose one or more pieces of PII about an individual. PII extraction attacks on LLMs can be categorized into two types: *targeted PII extraction* and *non-targeted PII extraction*, as outlined by [Chen *et al.*, 2024b; Zhang *et al.*, 2023]. Targeted PII extraction involves using partial PII already known to the attacker to guide the LLM in revealing additional PII about a specific individual. In contrast, non-targeted PII extraction leverages known PII of individuals within a specific profession to induce the LLM to disclose PII of as many individuals as possible in the same field.
- **PII Exposure via Fine-Tuning or User Prompts.** When a user fine-tunes LLMs with a dataset containing PII or inputs prompts that include PII, the model may memorize this information, leading to unintended leakage in subse-

quent queries. Additionally, the fine-tuning process itself can weaken the LLM's security mechanisms, exacerbating PII leakage from training data [Chen *et al.*, 2024b].

2.3 LLM Capabilities Contributing to PII Leakage

According to [Huang *et al.*, 2022], two capabilities of LLM could lead to privacy leakage: (i) Memorization: When an attacker provides the context of PII in the training data, LLMs can output the corresponding PII. (ii) Association: When an attacker queries an individual's information with providing their name, LLMs can infer and output the PII associated with that individual. [Carlini *et al.*, 2019] demonstrated that neural networks inevitably develop unintended memorization during training. This phenomenon refers to the unintended storage and potential leakage of rare or unique sequences from the training data, such as credit card numbers, social security numbers, or other private information. [Biderman *et al.*, 2024] studied the memorization capability of models and identified the phenomenon of emergent memorization, where models do not gradually memorize data but instead exhibit a sudden increase in memorization once their scale surpasses a certain threshold. [Shao *et al.*, 2024] investigated the association capabilities of LLMs. Their study demonstrated that PII could be extracted without requiring the exact prefix of the target information in the training data, highlighting vulnerabilities in the association mechanisms of LLMs.

3 Evaluation

In this section, we introduce datasets and evaluation metrics commonly used in existing research on PII leakage.

3.1 Datasets

The following datasets have been widely employed as test sets or ground truth for evaluating PII leakage.

- **Enron.** The Enron dataset[1] is a widely used public dataset comprising approximately 500,000 emails from Enron employees (2000–2002), including extensive PII such as names, email addresses, and phone numbers. Known for its large scale and rich metadata, it has been utilized to train models like Pythia and is frequently referenced in PII extraction research [Huang *et al.*, 2022; Li *et al.*, 2023; Nakka *et al.*, 2024a; Wang *et al.*, 2023].
- **WikiText.** Released by DeepMind in 2017, WikiText[2] contains over 100 million tokens from verified Wikipedia articles, including PII like names and birth dates. Studies such as [Borkar, 2023; Panda *et al.*, 2024; Chen *et al.*, 2024b] have used it for fine-tuning to examine PII leakage risks.
- **Pile.** The Pile[3], a large-scale 825GB text dataset from EleutherAI, aggregates sources such as books, research papers, articles, GitHub repositories, Wikipedia, social media, and blogs. Since Enron dataset is part of the Pile, it also contains significant PII, including names, email addresses, and phone numbers. It has been used to train open-source models (e.g., GPT-Neo) and validate PII extraction in [Zhang *et al.*, 2023; Shao *et al.*, 2024; Kim *et al.*, 2024].
- **Common Crawl.** The Common Crawl dataset[4] is a large-scale web archive regularly updated through web crawlers, gathering public webpages from across the globe. As of December 2024, Common Crawl includes 2.64 billion webpages (394TB). It contains diverse PII and serves as training data for models like GPT-3. Its subsets, including RefinedWeb[5], Dolma[6], and RedPajama[7], have been used in studies like [Nasr *et al.*, 2023] to verify PII authenticity.
- **Ai4Privacy.** Designed for privacy research, Ai4Privacy[8] is an open-source dataset containing annotated PII from social media, emails, online shopping, and healthcare records. It features PII-Masking-300k[9], specifically used for training and testing PII extraction in LLMs. Ai4Privacy has been employed in [Rashid *et al.*, 2025; Chen *et al.*, 2024b].
- **Industry Datasets.** Several industry-specific datasets are also used as test sets for PII extraction. For example, [Afreen *et al.*, 2020] uses the litigation case dataset ECHR[10] and the healthcare information dataset Yelp-Health[11], while [Li *et al.*, 2023] uses a self-built academic institution page dataset. [Rashid *et al.*, 2025] uses the news dataset MIND[12], and [Akkus *et al.*, 2024] uses the legal text dataset Freelaw[13], which is also a subset of the Pile. These datasets contain PII relevant to their respective professions, typically including personal information such as names, email addresses, physical addresses, and phone numbers.

3.2 Performance Metrics

The following metrics are commonly employed in studies on PII leakage.

- **TP and Accuracy.** The primary metrics are *True Positive (TP)* and *Accuracy*. TP represents the number of correctly extracted PII instances, while accuracy measures the proportion of correct extractions among all attempts. Accuracy is often represented as *Attack Success Rate (ASR)*, *Extraction Rate*, or *Hit Rate*. If extracting the same PII involves multiple attempts and at least one successful attempt is counted as TP, accuracy is often referred to as *Top-n Accuracy* or *hit@n* [Li *et al.*, 2023], where n denotes the number of execution attempts.
- **Recall and Precision.** When ground truth is unavailable for the extracted PII, additional verification is required. *Recall* measures the proportion of verifiably identified PII instances among all assessed instances and is often synonymous with ASR or extraction rate. When evaluating the effectiveness of the procedure in determining the validity of PII, *Precision* is used to measure the proportion of verifiable PII instances among all instances identified as PII.

[1] https://www.cs.cmu.edu/~enron/

[2] https://huggingface.co/datasets/Salesforce/wikitext

[3] https://pile.eleuther.ai/

[4] https://commoncrawl.org/

[5] https://huggingface.co/datasets/tiiuae/falcon-refinedweb

[6] https://allenai.org/dolma

[7] https://github.com/togethercomputer/RedPajama-Data

[8] https://www.ai4privacy.com/

[9] https://huggingface.co/datasets/ai4privacy/pii-masking-300k

[10] https://echr-opendata.eu/

[11] https://business.yelp.com/data/resources/open-dataset/

[12] https://msnews.github.io/

[13] https://github.com/thoppe/The-Pile-FreeLaw

- **Exact Match and Partial Match.** For certain types of PII that can be divided into multiple substructures, TP for PII extraction can be further categorized into *Exact Match* and *Partial Match* [Wang *et al.*, 2023]. Exact match refers to cases where the extracted PII is identical to the ground truth, while partial match indicates cases where the extracted PII shares some substructure (e.g., email domain or prefix) with the ground truth.
- **String Similarity.** For PII in string format without substructures, string similarity, such as *Cosine Similarity* and *Longest Common Subsequence (LCS)*, are used to measure extraction accuracy. Cosine similarity calculates the cosine of the angle between two strings treated as word frequency vectors, while LCS measures the length of the longest common subsequence between two strings [More *et al.*, 2024].
- **Other Metrics.** Additional metrics, such as *Attack Cost* and *Bypass Rate* [Chen *et al.*, 2024b], may also be utilized. Attack cost measures the financial overhead of PII extraction, while bypass rate quantifies the frequency of PII appearing in all LLM responses, indicating the effectiveness in circumventing the LLM's security alignment strategies.

4 Attack

In this section, we examine mainstream PII extraction attacks, categorized into three types: those originating from leaked training data, those executed through crafted prompts, and those conducted via fine-tuning, as detailed in §2.2. Table 1 provides a comparative analysis of representative studies on PII extraction from 2021 to 2024, focusing on core techniques, prompt formats, targeted victim LLMs, evaluation datasets, and the highest reported evaluation results. The common prompt formats for PII extraction (4th column) are summarized below, with a comprehensive review of these studies reserved for §4.1–§4.3.

Common Prompt Formats for PII Extraction. Prompts utilized in studies on PII extraction can be classified into five formats: (1) *Natural Language:* Direct queries, such as `What is the email of Jane Doe?`, designed to elicit PII from the LLM; (2) *Text Completion:* Incomplete passages, e.g., `Jane Doe [mailto:`, prompting the LLM to fill in missing PII, such as an email address; (3) *Template:* Structured prompts, e.g., `{"name": "Jane Doe", "email":`, formatted in JSON or SQL to query the LLM; (4) *True-Prefix:* Real textual prefixes from training data, where a prompt matching the prefix (e.g., `<prefix>`) increases the likelihood of the LLM generating the associated PII (e.g., `<PII>`); (5) *Few-Shot:* Providing the LLM with a small set of labeled examples to facilitate task understanding.

As shown in Table 1, the true-prefix method is mainly utilized in fine-tuned PII extraction tasks and often yields superior evaluation outcomes. The text completion approach is most frequently adopted for PII extraction via crafted prompts, whereas both true-prefix and natural language techniques are commonly applied for extracting PII from leaked training data. Furthermore, the few-shot techniques typically enhances evaluation results compared to scenarios where few-shot learning is absent, as evidenced by studies [Huang *et al.*, 2022; Chen *et al.*, 2024b].

4.1 Attack Type 1: Extracting PII from Leaked Training Data

This category of PII extraction attacks involves prompting the LLM to generate extensive training data, from which PII is extracted using methods like regular expressions. Early work [Carlini *et al.*, 2021] introduced a training data extraction attack that utilized black-box querying and model sampling techniques, extracting 78 PII samples (including names, addresses, and emails) from 1,800 candidate outputs.

Subsequent studies disrupt model alignment to exploit the LLM's memory capabilities. [Nasr *et al.*, 2023] developed a divergence attack, employing repetitive word generation prompts to deviate the model from its standard behavior. This approach resulted in 16.9% of 15,000 generated responses containing memorized PII, with 85.8% of it being authentic. Similarly, [Bai *et al.*, 2024] proposed an attack leveraging special characters in prompts to trigger the release of memorized PII, particularly phone numbers and emails. These methods highlight that inducing continuous, nonsensical outputs can inadvertently expose PII.

In contrast to general data leakage, [Zhang *et al.*, 2023] investigates targeted extraction, where attackers reconstruct missing content using partial knowledge. By optimizing model responses with a loss smoothing mechanism and employing local normalization for candidate suffixes, the study shows that longer prefixes improve PII extraction accuracy, while excessively long suffixes may impede reconstruction.

4.2 Attack Type 2: Extracting PII via Crafted Prompts

Recent research has focused on extracting PII from LLMs through carefully crafted prompts, utilizing memorization and association to enhance effectiveness. [Huang *et al.*, 2022] evaluated techniques such as direct querying, text completion, in-context learning, and few-shot learning to assess the impact of memorization and association on PII leakage, as discussed in §2.3. The study revealed that memorization poses a greater risk, as LLMs tend to recall and expose repeated data. While association is less effective for direct PII extraction, it facilitates inference and reconstruction when sufficient context is provided, with larger models exhibiting higher leakage risks. [Shao *et al.*, 2024] further quantified association capabilities, designing tasks to measure LLMs' ability to infer missing PII based on statistical patterns, confirming that larger models demonstrate stronger inference abilities.

Other studies explore inference techniques for PII extraction and verification. [Lukas *et al.*, 2023] introduced a game-based framework defining three PII leakage attacks: black-box extraction via blank queries, PII reconstruction from context, and PII inference (extending reconstruction by providing candidate options), achieving up to 10 times more PII leakage. [Niu *et al.*, 2023] investigated PII risks in code-generation models like GitHub Copilot, optimizing prompts with code-formatted prefixes to induce PII-containing completions. Using blind membership inference, the study found that 16.8 out of 1,000 Copilot queries resulted in PII leakage.

Emerging strategies focus on optimizing prompt formats for enhanced PII extraction. [Li *et al.*, 2023] proposed a

Attack Type	Representative Studies	Core Techniques	PII Extraction Prompt Format	Targeted Victim LLMs	Dataset	Highest Reported Evaluation Results
PII Extraction from Leaked Training Data	[Carlini *et al.*, 2021]	multi-strategy model sampling	true-prefix	GPT-2	-	78 Extracted PII of 1800 extracted data (TP)
	[Nasr *et al.*, 2023]	prompt-guided LLM misalignment	natural language	GPT-2, LLama-2 Falcon and *et al.*	Pile, RefinedWeb RedPajama, Dolma	2175 Extracted PII of 15000 extracted data (TP)
	[Bai *et al.*, 2024]	prompt with special characters	natural language	Falcon, GPT-3.5 LLama-2 and *et al.*	-	3.6% for GPT-3.5 (Accuracy for PII)
	[Zhang *et al.*, 2023]	loss smoothed soft prompting	true-prefix	GPT-Neo	Pile	62.8% (Recall for correct predicted suffix)
PII Extraction through Crafted Prompts	[Huang *et al.*, 2022]	measuring memorization and association	true-prefix, text completion, few -shot, template	GPT-Neo	Enron	37.06% (5-shot Accuracy for email)
	[Shao *et al.*, 2024]	measuring LLM's association capability	true-prefix template text completion	GPT-Neo, GPT-J-6B GPT-NeoX	Pile	3.31% for GPT-NEOX (Accuracy for email)
	[Lukas *et al.*, 2023]	PII reconstruction and inference	text completion	GPT-2	ECHR, Enron Yelp-Health	70% in ECHR (Accuracy for PII)
	[Niu *et al.*, 2023]	membership inference for PII validation	template	Codex (GPT-3)	-	1.68% (top-5 Accuracy for PII)
	[Li *et al.*, 2023]	COT prompting for PII extraction	natural language text completion	GPT-3.5	Enron, Academic Institution Page	59.09% in Enron (Accuracy for email)
	[Nakka *et al.*, 2024a]	adding other PII as true prefix	true-prefix text completion	GPT-J-6B	Enron	6.86% (top-2308 Accuracy for phone)
	[Kim *et al.*, 2024]	soft prompt tuning	text completion	OPT-1.3B	Pile	1.3% (Accuracy for phone)
PII Extraction through Fine-tuning	[Chen *et al.*, 2024b]	PII recovery via fine-tuning	text completion few-shot	GPT-2, GPT-3.5 Llama-2 and *et al.*	Enron, Ai4Privacy ECHR, WikiText	69.90% for GPT-3.5 in Enron (5-shot Accuracy for email)
	[Akkus *et al.*, 2024]	fine-tuning with LLM-generated data	true-prefix, text completion, few -shot, template	Pythia Suite Llama-3 and *et al.*	Enron, Freelaw	58 extracted email in Enron for Pythia-2.8b (TP)
	[Panda *et al.*, 2024]	adding posion prefix via fine-tuning	true-prefix	Pythia Suite	Enron, WikiText	50% in Enron (PII accuracy)
	[Borkar, 2023]	PII extraction after fine-tuning	true-prefix	GPT-2	Enron, WikiText-103 Common Crawl	44 Extracted email in Enron (TP)
	[Rashid *et al.*, 2025]	bounded unlearning for posioning	true-prefix	Llama2-7B GPT-Neo-1.3B	MIND, WikiText-103 Ai4Privacy	177 Extracted PII of 500 extracted data in MIND for LLama2-7B(TP)

Table 1: Review of mainstream PII leakage attacks across three categories, published between 2021 and 2024

multi-step jailbreaking attack inspired by Chain-of-Thought (CoT), dividing the process into sub-tasks to bypass security alignment, achieving a 59.09% extraction rate for frequent Enron email addresses on GPT-3.5. [Nakka *et al.*, 2024a] demonstrated that adding unrelated PII as a prefix improved extraction rates by 5 to 18 times, particularly for phone numbers. [Kim *et al.*, 2024] extended research into white-box attacks, using soft prompt tuning to refine black-box prompts, increasing extraction success rates from 0.0047% to 1.3%.

4.3 Attack Type 3: Extracting PII via Fine-tuning

Fine-tuning adapts pre-trained models to specific tasks but significantly heightens PII leakage risks. [Chen *et al.*, 2024b] fine-tuned LLMs on a small PII-containing dataset until its perplexity fell below a threshold, reinforcing memorization. Notably, fine-tuning could also restore forgotten PII from the original training data. Experiments on targeted and non-targeted PII extraction revealed that fine-tuning APIs are highly susceptible, and conventional privacy mechanisms fail to counter this attack. In contrast, [Akkus *et al.*, 2024] investigated fine-tuning on LLM-generated PII datasets. Using Pythia fine-tuned on Enron to produce synthetic PII, they fine-tuned another LLM and observed that it still leaked real PII, despite training on seemingly unrelated synthetic data. [Panda *et al.*, 2024] introduced a neural phishing attack, embedding a poison prefix into the fine-tuning dataset. After training, attackers with partial prior knowledge and the poison prefix could extract PII, particularly numeric data like phone numbers. Even if a PII instance appeared only once, the model memorized and later reproduced it.

Research has explored unlearning to mitigate PII leakage induced by fine-tuning. [Borkar, 2023] found that removing easily extractable data increased the vulnerability of new data. Using a fine-tuned GPT-2 model trained on the WikiText-103 dataset, they showed that fine-tuned LLMs leak PII not only from fine-tuning but also from pre-training. [Rashid *et al.*, 2025] proposed bounded unlearning, a poisoning technique that amplifies privacy leakage by maximizing loss on noise data. Their evaluation showed that poisoned fine-tuned models were more prone to membership inference and PII extraction attacks, making it easier to identify training datasets. Despite amplifying PII leakage, fine-tuning did not significantly impair overall model performance.

4.4 Evaluation Benchmarks for PII Extraction

Several studies have proposed benchmarks to evaluate PII leakage. [Nakka *et al.*, 2024b] introduced PII-Scope, a comprehensive benchmark assessing LLMs' PII leakage across various privacy attack scenarios, including true-prefix [Carlini *et al.*, 2021; Carlini *et al.*, 2022], template [Huang *et al.*, 2022], few-shot [Huang *et al.*, 2022], PII-Compass [Nakka *et al.*, 2024a], and soft prompt embeddings attacks [Kim *et al.*, 2024]. Evaluations under single-query, multi-query, and fine-tuning conditions reveal that existing PII attacks can triple extraction efficiency with limited query budgets. [Wang *et al.*, 2023] evaluated GPT-3.5 and GPT-4 for PII leakage using Enron, targeting email addresses. Employing techniques such as direct queries, code completion, and few-shot prompting, the study found that GPT-4 and GPT-3.5 achieved PII extraction rates of 48.19% and 44.47%, respectively.

[Li *et al.*, 2024] proposed LLM-PBE, a privacy evaluation framework encompassing multiple attacks, including data extraction, membership inference, jailbreaking, and prompt leaking. It assesses privacy risks across LLM's lifecycle, from training to deployment. Experiments on Llama-2 demonstrated that data type, length, and pretraining size influence privacy risks, with richer contextual PII more likely to be memorized during fine-tuning. The study also highlighted that training data containing sensitive personal or domain-specific information increases the risk of unintended PII leakage. [More *et al.*, 2024] investigated extraction attacks from an adversarial perspective, integrating multiple attack strategies and leveraging defense-unprocessed model checkpoints. The study explored how model size and prompt variations affect attack performance, incorporating LLM-PBE [Li *et al.*, 2024] to measure extraction risks across model checkpoints and sizes. Results indicated that access to multiple model checkpoints significantly increases PII extraction rates.

4.5 Comparison between PII Extraction Attacks

PII extraction based on leaked training data benefits from a simple attack setup, requiring only natural language or true-prefix prompts to retrieve large amounts of training data, even from smaller models like GPT-2 and GPT-Neo. For instance, [Zhang *et al.*, 2023] achieved up to 62.8% recall in PII extraction using a predicted suffix verbatim approach. However, a major limitation is that only a small portion of the retrieved training data contains PII.

In contrast, PII extraction via crafted prompts employs diverse prompt formats and targets LLMs of varying parameter sizes, closely mirroring real-world attack scenarios. Studies such as [Lukas *et al.*, 2023; Li *et al.*, 2023] report high success rates in targeted PII extraction. Nevertheless, its generalizability is questionable, as evaluations often rely on limited datasets and define success based on extracting PII even once across multiple attempts, inflating reported accuracy.

Fine-tuning-based PII extraction can retrieve PII not only from fine-tuning datasets but also from the original training data, often with high accuracy. However, it requires true-prefix prompts for effective extraction, limiting its applicability when the fine-tuning or training dataset is unknown. Additionally, this method becomes ineffective when access to fine-tuning interfaces is restricted.

5 Defense

The core defense techniques against PII extraction attacks can be categorized as follows:

- **Defense during Model Training.** This approach addresses sensitive information during the training phase. Common techniques include data cleaning [Kandpal *et al.*, 2022] and differential privacy [Hoory *et al.*, 2021]. While iterative data cleaning enhances privacy protection, it requires retraining the model, which is computationally expensive.
- **Defense via Model Adaptation.** This approach modifies parameters or structures of a trained LLM. Techniques include model editing [Meng *et al.*, 2022], fine-tuning [Yu *et al.*, 2021], and unlearning [Jang *et al.*, 2022]. Although cost-effective by reducing training costs, residual memorization of sensitive data may persist.
- **Defense during Query Execution.** This approach alters user queries (e.g., by modifying or filtering prompts) during LLM execution to prevent sensitive data exposure.

Since most publicly available LLMs are pre-trained models, retraining is often impractical for users, making post-training defenses through model adaptation or query-time interventions more prevalent. Table 2 reviews recent studies on post-training defenses against PII leakage.

Defenses	Representative Studies	Core Techniques	Involved LLMs
Defense via Model Adaptation	[Chen *et al.*, 2024a]	deactivate neurons storing PII	GPT-Neo
	[Wu *et al.*, 2023]	privacy attribution score ranking	BERT
	[Ashuach *et al.*, 2024]	privacy rank editing for neuron	GPT-J-6B
	[Venditti *et al.*, 2024]	private association editing for neuron	GPT-J-6B
Defense during Query Execution	[Zhou *et al.*, 2024]	user-led data minimization	Llama3-8B GPT-4o
	[Sun *et al.*, 2024]	generative desensitization	BERT, GPT-3.5 Llama-3 and *et al.*
	[Frikha *et al.*, 2024]	private attribute randomization	Llama3 Phi-3
	[Deshmukh *et al.*, 2023]	PII obfuscation	GPT-3.5
	[Chong *et al.*, 2024]	local LLM-based topic identification	GPT-3.5
	[Siyan *et al.*, 2024]	local LLM privacy filtering	GPT-4o-mini Llama-3 and *et al.*

Table 2: Review of Mainstream PII Leakage Defenses

5.1 Defense via Model Adaptation

This category of defenses focuses on PII unlearning and editing PII-related neurons in LLMs. [Chen *et al.*, 2024a] employs learnable binary weight masks to identify and deactivate neurons storing PII. Experiments reveal that most privacy neurons are located in MLP layers, and deactivating them reduces PII memorization accuracy on the Enron dataset from 45.83% to 5.60%. Similarly, [Wu *et al.*, 2023] uses gradient attribution to assign privacy scores to neurons, identifying high-risk neurons and erasing their activation values to prevent PII retention. The study finds that privacy neurons are concentrated in the upper Transformer layers and aggregate over time during training. Inspired by [Wu *et al.*, 2023], [Ashuach *et al.*, 2024] proposes REVS (Rank Editing in the Vocabulary Space), which identifies sensitive tokens, maps them into the vocabulary space to locate critical layers and neurons influencing token generation, and lowers their ranking in the model's output. This approach reduces memorization while maintaining general text generation capabilities, achieving 99.95% deletion accuracy on SSNs and 97.22% on emails. [Venditti *et al.*, 2024] introduces Private Association Editing (PAE) to modify key neurons and disrupt associations between PII and its owner. By defining PAE rules for data removal and replacement and adjusting key parameters in Feed-Forward Networks, PAE reduces privacy leakage by 60.52% under a 200-token prompt attack on the Enron dataset.

5.2 Defense during Query Execution

Defenses at query time primarily focus on detecting and replacing user PII before it is processed by a LLM.

PII Replacement. [Zhou *et al.*, 2024] introduces a user-led data minimization approach, developing Rescriber, a browser extension that enables users to proactively detect and replace PII with placeholders or general expressions. Supporting both Llama3-8B and GPT-4o, it achieves 0.74 precision and 0.87 recall for PII detection in GPT-4o. [Sun *et al.*, 2024] proposes a prompt-level privacy framework combining fine-tuned PII identification (95.95% accuracy) with adversarial desensitization, replacing PII with misleading or generalized data. It also incorporates adversarial perturbation by injecting special symbols, complicating PII reconstruction while maintaining model comprehension. [Frikha *et al.*, 2024] develops IncogniText, which uses Private Attribute Randomization (PAR) to replace real attributes with multiple plausible fabricated ones. Evaluations on Llama-based models show a reduction in correctly predicted private attributes from 71.2% to 15.4%. [Deshmukh *et al.*, 2023] introduces a Transformer-based PII obfuscation framework that replaces sensitive data with Faux-PII while preserving usability. Its API automatically obfuscates input before LLM processing and restores transformed data in the output, leveraging user provided tokens, NER, and part-of-speech substitution to balance privacy and utility.

Locally Deployed LLMs for Privacy. [Chong *et al.*, 2024] presents Casper, a client-side privacy filter that removes PII before sending queries to LLMs. It employs a three-layer filtering system: Rule-Based, ML-Based NER, and Local LLM Topic Identification, achieving 98.5% PII detection accuracy and 89.9% topic detection accuracy on 4,000 synthetic queries. [Siyan *et al.*, 2024] proposes PAPILLON, a hybrid approach combining local and remote LLMs. Upon receiving a user's prompt, a local LLM filters the input to remove or transform PII before forwarding partially desensitized queries to a remote API-based LLM when necessary. Compared to GPT-4o-mini, PAPILLON reduces privacy leakage from 100% to 7.5%, with only a slight response quality drop from 88.2% to 85.5%.

6 Challenges and Future Directions

In this section, we discuss unresolved issues in PII leakage and propose future research directions.

Evaluation of PII Extraction Methods on Diverse Datasets. As shown in Table 1, many studies rely heavily on the Enron dataset, particularly those achieving high attack efficacy, such as [Chen *et al.*, 2024b; Li *et al.*, 2023], whose extraction accuracy is evaluated exclusively on Enron. However, the effectiveness of these techniques on other datasets remains unclear. Additionally, the Enron's PII is limited to specific professions, raising concerns about the generalizability of extraction methods. Future research should focus on constructing large-scale PII datasets encompassing a wider range of professions to enable comprehensive evaluations.

Increased Focus on Non-Targeted PII Extraction. As categorized in §2.2, most PII extraction techniques in §4 are targeted, with only a few studies exploring non-targeted approaches, such as training data leakage-based methods [Carlini *et al.*, 2021; Lukas *et al.*, 2023] and fine-tuning-based methods like [Chen *et al.*, 2024b]. Given that non-targeted extraction is better suited for retrieving extensive PII from LLMs, future research should prioritize developing and evaluating non-targeted PII extraction techniques to better assess the scale and severity of PII leakage.

Comprehensive Evaluation and Benchmarking of PII Leakage. While prior research often evaluates individual PII extraction methods, broader security assessments considering multiple privacy leakage techniques are limited, with notable efforts in [Nakka *et al.*, 2024b; Li *et al.*, 2024]. Additionally, there is a lack of standardized evaluation datasets or comprehensive PII query prompt benchmarks. Future work should establish an extensive PII leakage evaluation framework and benchmark to enhance LLM privacy protections and enable thorough assessments of real-world PII leakage threats.

Malicious User Prompt Detection. Current PII leakage defenses primarily focus on detecting and replacing user-inputted PII but rarely analyze the malicious intent behind prompts. [Chong *et al.*, 2024] explores using local LLMs to analyze prompt topics. Future research should leverage NLP techniques to assess prompt intent, detect PII in LLM outputs, and develop defenses against extraction attacks. Capturing and analyzing wild PII extraction prompts could also improve proactive detection mechanisms.

Semantic-Level PII Extraction and Defense. Most existing PII extraction and defense techniques rely on direct matching of PII entities. However, if PII is represented in obfuscated natural language or indirect hints, new privacy risks may arise. Future research should explore context-aware PII identification and LLM-based semantic parsing to develop novel attack and defense mechanisms capable of detecting and mitigating indirect PII leakage.

7 Conclusion

In this paper, we present a comprehensive survey on PII leakage in LLMs. We begin by defining and categorizing PII leakage, followed by a systematic review of datasets and evaluation metrics commonly used. We then categorize and analyze attack and defense techniques, outlining the current research landscape. Finally, we discuss existing challenges, propose future research directions, and provide insights to guide further exploration.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (No.2023YFB3106800) and the National Natural Science Foundation of China (No.62272410).

References

[Afreen *et al.*, 2020] Asad Afreen, Moosa Aslam, and Saad Ahmed. Analysis of fileless malware and its evasive behavior. In *ICCWS*, pages 1–8. IEEE, 2020.

[Akkus *et al.*, 2024] Atilla Akkus, Mingjie Li, Junjie Chu, Michael Backes, Yang Zhang, and Sinem Sav. Generated data with fake privacy: Hidden dangers of fine-tuning large language models on generated data. *arXiv:2409.11423*, 2024.

[Ashuach *et al.*, 2024] Tomer Ashuach, Martin Tutek, and Yonatan Belinkov. Revs: Unlearning sensitive information in language models via rank editing in the vocabulary space. *arXiv:2406.09325*, 2024.

[Bai *et al.*, 2024] Yang Bai, Ge Pei, Jindong Gu, Yong Yang, and Xingjun Ma. Special characters attack: Toward scalable training data extraction from large language models. *arXiv:2405.05990*, 2024.

[Biderman *et al.*, 2024] Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *NeurIPS*, 36, 2024.

[Borkar, 2023] Jaydeep Borkar. What can we learn from data leakage and unlearning for law? *arXiv:2307.10476*, 2023.

[Carlini *et al.*, 2019] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security*, pages 267–284, 2019.

[Carlini *et al.*, 2021] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, et al. Extracting training data from large language models. In *USENIX Security*, pages 2633–2650, 2021.

[Carlini *et al.*, 2022] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv:2202.07646*, 2022.

[Carlini *et al.*, 2023] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security*, pages 5253–5270, 2023.

[Chen *et al.*, 2024a] Ruizhe Chen, Tianxiang Hu, Yang Feng, and Zuozhu Liu. Learnable privacy neurons localization in language models. *arXiv:2405.10989*, 2024.

[Chen *et al.*, 2024b] Xiaoyi Chen, Siyuan Tang, Rui Zhu, Shijun Yan, Lei Jin, Zihao Wang, Liya Su, Zhikun Zhang, XiaoFeng Wang, and Haixu Tang. The janus interface: How fine-tuning in large language models amplifies the privacy risks. In *ACM CCS*, pages 1285–1299, 2024.

[Chong *et al.*, 2024] Chun Jie Chong, Chenxi Hou, Zhihao Yao, and Seyed Mohammadjavad Seyed Talebi. Casper: Prompt sanitization for protecting user privacy in web-based large language models. *arXiv:2408.07004*, 2024.

[Deng *et al.*, 2024] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Masterkey: Automated jailbreaking of large language model chatbots. In *ISOC NDSS*, 2024.

[Deshmukh *et al.*, 2023] Ajinkya Deshmukh, Saumya Banthia, and Anantha Sharma. Life of pii–a pii obfuscation transformer. *arXiv:2305.09550*, 2023.

[Frikha *et al.*, 2024] Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. Incognitext: Privacy-enhancing conditional text anonymization via llm-based private attribute randomization. *arXiv:2407.02956*, 2024.

[Hoory *et al.*, 2021] Shlomo Hoory, Amir Feder, Avichai Tendler, Sofia Erell, Alon Peled-Cohen, Itay Laish, Hootan Nakhost, Uri Stemmer, et al. Learning and evaluating a differentially private pre-trained language model. In *EMNLP*, pages 1178–1189, 2021.

[Huang *et al.*, 2022] Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In *EMNLP*, pages 2038–2047, December 2022.

[Jang *et al.*, 2022] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv:2210.01504*, 2022.

[Kandpal *et al.*, 2022] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR, 2022.

[Kim *et al.*, 2024] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. Propile: Probing privacy leakage in large language models. *NeurIPS*, 36, 2024.

[Li *et al.*, 2023] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. In *EMNLP*, pages 4138–4153, 2023.

[Li *et al.*, 2024] Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, et al. Llm-pbe: Assessing data privacy in large language models. *arXiv:2408.12787*, 2024.

[Lukas *et al.*, 2023] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE, 2023.

[McCallister *et al.*, 2010] Erika McCallister, Timothy Grance, and Karen A Scarfone. Sp 800-122. guide to protecting the confidentiality of personally identifiable information (pii), 2010.

[Meng *et al.*, 2022] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv:2210.07229*, 2022.

[More *et al.*, 2024] Yash More, Prakhar Ganesh, and Golnoosh Farnadi. Towards more realistic extraction attacks: An adversarial perspective. *arXiv:2407.02596*, 2024.

[Nakka *et al.*, 2024a] Krishna Kanth Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. Pii-compass: Guiding llm training data extraction prompts towards the target pii via grounding. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 63–73, 2024.

[Nakka *et al.*, 2024b] Krishna Kanth Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. Pii-scope: A benchmark for training data pii leakage assessment in llms. *arXiv:2410.06704*, 2024.

[Nasr *et al.*, 2023] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv:2311.17035*, 2023.

[Niu *et al.*, 2023] Liang Niu, Shujaat Mirza, Zayd Maradni, and Christina Pöpper. {CodexLeaks}: Privacy leaks from code generation language models in {GitHub} copilot. In *USENIX Security*, 2023.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

[Panda *et al.*, 2024] Ashwinee Panda, Christopher A. Choquette-Choo, Zhengming Zhang, Yaoqing Yang, and Prateek Mittal. Teach llms to phish: Stealing private information from language models. In *ICLR*, 2024.

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

[Rashid *et al.*, 2025] Md Rafi Ur Rashid, Jing Liu, Toshiaki Koike-Akino, Ye Wang, and Shagufta Mehnaz. Forget to flourish: Leveraging machine-unlearning on pretrained language models for privacy leakage. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39, 2025.

[Shao *et al.*, 2024] Hanyin Shao, Jie Huang, Shen Zheng, and Kevin Chen-Chuan Chang. Quantifying association capabilities of large language models and its implications on privacy leakage. In *EACL*, pages 814–825, 2024.

[Shen *et al.*, 2024] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now"': Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *ACM CCS*, 2024.

[Siyan *et al.*, 2024] Li Siyan, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. Papillon: Privacy preservation from internet-based and local language model ensembles. *arXiv preprint arXiv:2410.17127*, 2024.

[Staab *et al.*, 2024] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *ICLR*, 2024.

[Sun *et al.*, 2024] Xiongtao Sun, Gan Liu, Zhipeng He, Hui Li, and Xiaoguang Li. Deprompt: Desensitization and evaluation of personal identifiable information in large language model prompts. *arXiv:2408.08930*, 2024.

[Venditti *et al.*, 2024] Davide Venditti, Elena Sofia Ruzzetti, Giancarlo A Xompero, Cristina Giannone, Andrea Favalli, Raniero Romagnoli, and Fabio Massimo Zanzotto. Enhancing data privacy in large language models through private association editing. *arXiv:2406.18221*, 2024.

[Wang *et al.*, 2023] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.

[Wu *et al.*, 2023] Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. Depn: Detecting and editing privacy neurons in pretrained language models. *arXiv:2310.20138*, 2023.

[Yan *et al.*, 2024] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. On protecting the data privacy of large language models (llms): A survey. *arXiv:2403.05156*, 2024.

[Yao *et al.*, 2024] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.

[Yu *et al.*, 2021] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, et al. Differentially private fine-tuning of language models. *arXiv:2110.06500*, 2021.

[Yu *et al.*, 2024] Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. {LLM-Fuzzer}: Scaling assessment of large language model jailbreaks. In *USENIX Security*, pages 4657–4674, 2024.

[Zhang *et al.*, 2023] Zhexin Zhang, Jiaxin Wen, and Minlie Huang. Ethicist: Targeted training data extraction through loss smoothed soft prompting and calibrated confidence estimation. In *ACL*, pages 12674–12687, 2023.

[Zhou *et al.*, 2024] Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. Rescriber: Smaller-llm-powered user-led data minimization for navigating privacy trade-offs in llm-based conversational agent. *arXiv:2410.11876*, 2024.

[Zou *et al.*, 2023] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv:2307.15043*, 2023.