# Beyond Data Privacy: New Privacy Risks for Large Language Models

Yuntao Du† , Zitao Li‡, Ninghui Li†, Bolin Ding‡
† Department of Computer Science, Purdue University
{ytdu, ninghui}@purdue.edu
‡ Alibaba
{zitao.l, bolin.ding}@alibaba-inc.com

**Abstract**

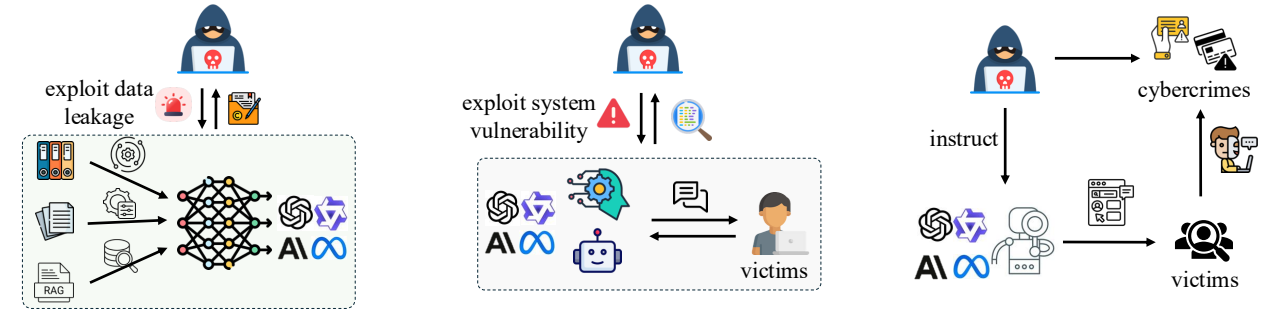
Large Language Models (LLMs) have achieved remarkable progress in natural language understanding, reasoning, and autonomous decision-making. However, these advancements have also come with significant privacy concerns. While significant research has focused on mitigating the data privacy risks of LLMs during various stages of model training, less attention has been paid to new threats emerging from their deployment. The integration of LLMs into widely used applications and the weaponization of their autonomous abilities have created new privacy vulnerabilities. These vulnerabilities provide opportunities for both inadvertent data leakage and malicious exfiltration from LLM-powered systems. Additionally, adversaries can exploit these systems to launch sophisticated, large-scale privacy attacks, threatening not only individual privacy but also financial security and societal trust. In this paper, we systematically examine these emerging privacy risks of LLMs. We also discuss potential mitigation strategies and call for the research community to broaden its focus beyond data privacy risks, developing new defenses to address the evolving threats posed by increasingly powerful LLMs and LLM-powered systems.

## 1 Introduction

Recent advancements in deep learning, particularly in natural language processing, have led to the development of large language models (LLMs). Over the past few years, LLMs have demonstrated impressive capabilities in understanding and generating human language. These models are rapidly growing in size and effectiveness, yielding breakthroughs and attracting increasing research and social attention. Beyond natural language understanding, their emergent abilities [173] have enabled them to achieve unparalleled performance on complex tasks. As a result, LLMs are no longer standalone models but are increasingly integrated as core decision-making components in larger systems, such as interactive chatbots [14, 127, 186] and autonomous agents [132, 144, 187].

However, this rapid development comes with growing concerns about its privacy implications. As a primary source of privacy risk, LLMs are trained on vast, internet-scale corpora that often contain sensitive personal information and copyrighted content. These data privacy risks are amplified when models are fine-tuned on private, proprietary datasets. Studies [32, 107, 109] have shown that LLMs can memorize and inadvertently leak training data across various model learning stages, raising issues related to training data extraction [25], copyright infringement [172], and test set contamination [129].

Beyond the risks of training data leakage, privacy threats also emerge from the integration of LLMs into larger, more complex systems, which we refer to as LLM-powered systems. These systems, especially those that use LLMs as decision-making engines in agent-based applications [36, 169], introduce new vulnerabilities and expand the potential attack surface for privacy violations. For instance, a user may share personal information with an LLM-based chatbot in order to receive personalized responses or




(a) Data privacy risks of various learning stages (i.e., pre-training, fine-tuning, and in-context learning).

(b) Privacy risks in LLM-powered systems, such as side-channel and information exfiltration.

(c) Privacy risks from malicious use of LLMs, such as automated profiling and social engineering.

Figure 1: Illustrations of different types of privacy risks posed by large language models.

suggestions. However, this information could be exfiltrated through side channels [23] or unintentionally disclosed by the model itself [158]. Such risks are not inherent to the LLM alone but emerge from the architecture of interactions between users, models, and other system components. As LLM-powered applications become increasingly widespread in both daily life and professional domains, these privacy risks become more prominent and urgent.

A third category of privacy threats arises from the advanced reasoning and autonomous decision-making capabilities of LLMs, which create new opportunities for malicious exploitation. These capabilities enable adversaries to automate sophisticated attacks at unprecedented scale and speed, substantially lowering the barrier to entry for cyberattacks. For instance, an attacker could instruct an LLM to infer sensitive attributes, such as a user's demographics, from their public online posts, leading to de-anonymization and other severe cybercrimes [49, 101]. Similarly, LLMs can be leveraged to launch large-scale, highly personalized social engineering campaigns, resulting in significant financial and societal consequences [106].

Positioning and Contribution Existing research has predominantly focused on training data privacy issues of LLMs. In contrast, less attention has been paid to the privacy threats posed by LLM-powered systems and the malicious use of LLMs. The privacy risks arising from LLM-powered systems and their malicious use represent a paradigm shift. These risks are not rooted in sensitive training data, but in the increasingly powerful autonomy of LLMs. As a result, existing data privacy frameworks may not always be well-suited to analyze or mitigate these emerging threats. This paper aims to bridge this gap by providing a comprehensive study of the new threat landscape introduced by LLMs. We systematically examine the privacy risks that arise from these models and their applications, and discuss potential mitigation strategies, calling for research efforts and greater public awareness to address these emerging privacy challenges.

Related Work Several studies [48, 94, 96, 99, 153, 176] have surveyed the data privacy risks of LLMs and explored mitigation strategies, However, these studies do not cover the emerging privacy threats posed by the increasing integration of LLM-powered systems and the potential malicious use of LLMs, which we identify as critical new privacy risks. Another line of research [18, 30, 97, 180] has examined the privacy implications of LLMs during user interactions. Our work builds on and extends these prior studies, providing a systematic and comprehensive analysis of the privacy risks posed by LLMs.

Roadmap In Section 2, we introduce the necessary background on LLMs and their key developmental trends. In Section 3, we discuss the primary data privacy risks associated with LLMs at different stages of training. In Section 4, we analyze the privacy threats of LLM-powered systems and discuss potential mitigations. In Section 5, we examine the privacy threats arising from the malicious use of LLMs.

Finally, we conclude and discuss future directions in Section 8.

## 2 Background on Large Language Models

A language model (LM) is a type of machine learning model for natural language processing. In general, an LM estimates the generative likelihood of sequences of words by predicting the probabilities of future or missing tokens[1]. In recent years, large language models (LLMs), trained on massive datasets for token prediction, have achieved unprecedented performance across a wide range of applications [205]. The scaling of these models has also unlocked remarkable emergent abilities not present in their smaller counterparts [173], including in-context learning [45], analogical reasoning [171], and the capacity to power autonomous agents [44].

With the rapid development of LLMs, LLM agents have emerged and become increasingly popular. These are intelligent entities powered by LLMs that are capable of autonomously carrying out complex tasks—such as conducting in-depth research or managing computer operations—while adapting to specific user needs [169]. This shift represents not only a major technological advancement but also a reimagining of human–machine interaction. Their impressive capabilities have already been applied in a wide range of domains, from chatbots [1] to professional tools like programming assistants [36].

Trends on LLMs Despite these advancements, the rapid evolution of LLMs also introduces new privacy risks. We identify three key trends that are closely related to growing privacy challenges:

- *Trained on Sensitive Data.* LLMs are trained on vast amounts of diverse data, which may include sensitive or copyrighted information. Moreover, the advanced capabilities of these models also enable them to access and utilize sensitive data through fine-tuning or in-context learning, especially when building proprietary models or personalized LLM-based applications [203]. This further increases the potential privacy risk of sensitive data that is trained on LLMs.

- *Incorporating into Popular Applications.* LLMs are increasingly embedded into core components of widely used software and professional tools. For example, they serve critical roles in domains such as code generation for software development, document analysis in legal and medical contexts, and as reasoning engines for autonomous agents that interact with external systems and platforms. As LLMs become ubiquitous in these daily applications, the surface area for privacy risks expands significantly, as adversaries can exploit these models in a wide range of sensitive contexts.

- *Growing Capability and Accessibility.* LLMs are rapidly evolving beyond text-only capabilities. Recent advances in vision-language models (VLMs) enable multimodal reasoning over both text and images [128, 162]. At the same time, access to powerful commercial and open-source models has become dramatically easier and much more affordable [108]. This combination of greater capability and accessibility empowers adversaries, giving them opportunities to exploit these advanced, autonomous systems to perform privacy-infringing attacks.

Taken together, these trends create a new privacy landscape that both amplifies data privacy concerns and introduces new privacy threats. The following sections will analyze these risks in detail, exploring their implications and potential mitigation strategies.

## 3 Data Privacy Risks in LLMs

LLMs themselves pose a significant privacy risk, as they, like other machine learning models, have been shown to memorize elements of their training data [22, 25]. These data privacy risks can arise at multiple

[1]A token refers to the smallest semantic unit processed by the model, which can be a character, subword, or word.

stages of the model learning process. From a parametric training perspective, LLMs typically undergo both a large-scale *pre-training stage* and subsequent *fine-tuning stages*. While LLMs are pre-trained and fine-tuned on massive corpora, and most state-of-the-art systems [14, 19, 53, 127, 186] do not disclose the provenance of their training data, concerns regarding potential privacy breaches have become increasingly pronounced. Moreover, due to their generative nature, LLMs support *in-context* learning, which provides a simple yet powerful mechanism for adapting inference behavior without modifying model parameters. These unique characteristics not only enhance the utility of LLMs but also broaden the avenues through which sensitive information may be exposed. In the following discussion, we examine two of the most prominent privacy threats to LLMs, membership inference and training data extraction, along with their privacy implications and potential enhancements.

### 3.1 Membership Inference Attack

Closely connected with Differential Privacy (DP) [54, 95], Membership Inference Attack (MIA) [152] has become a widely adopted approach for privacy auditing of machine learning (ML) models [119, 157]. MIA assesses how much a trained ML model reveals about its training data by determining whether specific query instances (or documents) were included in the dataset. In the context of LLMs, membership inference attacks have been applied at different stages to explore the privacy risk.

- *Pre-training Stage.* Early studies [110, 150, 181, 201] propose various membership signals that are derived from LLM's outputs to distinguish members from non-members. However, subsequent studies [111, 197] have identified fundamental flaws in the evaluation methodologies of prior studies. Particularly, the use of temporal data to separate members from non-members introduces subtle distributional shifts between the two groups, resulting in unreliable attack performance that does not accurately reflect true privacy leakage. Under more rigorous setups, existing MIAs [27, 37, 83, 198] perform no better than random guessing when targeting pre-training data of LLMs. This ineffectiveness is largely due to the fact that each data point is typically used only once during pre-training [118, 166], and the vast diversity of training corpora further dilutes the influence of a single example [51, 68].

- *Fine-tuned Stage.* Following the pre-training stage, the fine-tuning stage requires substantially fewer resources and focuses on adapting pre-trained models to domain-specific downstream tasks. However, fine-tuning datasets frequently contain personally identifiable information (PII) [32], copyrighted material [103], or sensitive organizational data [11]. MIAs against fine-tuned LLMs leverage techniques such as prompt calibration [61], hypothesis testing [113], and ensemble methods [199]. Moreover, recent work [58] has shown that human preference data used for alignment tuning (via Direct Preference Optimization (DPO) [185]) is also vulnerable to MIAs. Compared to attacks on pre-trained LLMs, MIAs against fine-tuned models are markedly more effective. This increased vulnerability arises because fine-tuning datasets are considerably smaller, and fine-tuning often involves multiple training epochs. These factors increase the model's memorization of sensitive data, posing heightened privacy threats in fine-tuned LLMs.

- *In-context Learning Stage.* Fine-tuning LLMs for specific domains involves non-trivial computation via parameter updates. In-Context Learning (ICL) [45] has emerged as a popular, more efficient adaptation paradigm, as it does not require modifying model parameters. In ICL, private data are provided as demonstrations within the prompt itself to guide the model's inference for a specific task. These demonstrations can be manually prepared [19] or dynamically retrieved from a private knowledge base using Retrieval-Augmented Generation (RAG) systems [15, 93, 151]. The vulnerability of ICL to MIAs has been explored using methods such as prompt injection [13],

analysis of semantic similarity [59], or measuring contextual influence [59]. Since ICL relies on only a few demonstration examples, each one has a significant impact on the model's output and performance, making these attacks highly effective at identifying private data.

Growing Threats While current MIAs have been effective at assessing privacy risks for fine-tuned and in-context data in LLMs, they have shown limited success against pre-trained models. However, this does not mean that pre-trained models are free from privacy risks. Two emerging research directions make MIAs more powerful, presenting a growing threat to the privacy of pre-training data. The first direction shifts the focus from analyzing individual data instances to examining larger collections of data. Recent studies [107] show that instead of detecting membership at the sentence level, aggregating membership signals across multiple sentences within a document can reliably reveal whether that dataset was included in training. This has been extended to paragraphs and entire collections, showing that these larger structures are also vulnerable to MIAs [138]. The second direction involves more advanced and computationally intensive attacks. Recent research [65] shows that training hundreds of shadow models [152] to exploit behavioral discrepancies can significantly enhance MIA effectiveness. These stronger attacks, building on prior successes against classifiers [21, 194], further elevate the risks to pre-trained data privacy. Together, these emerging research trends indicate that future improvements in MIAs could amplify the privacy risks of pre-trained LLMs.

Privacy Implications Effective MIAs on LLMs have serious trustworthy risks, such as the leakage of copyrighted content [52, 172] or test set contamination from evaluation benchmarks [129]. Moreover, MIAs can serve as a fundamental building block for more sophisticated attacks, such as training data extraction [25, 26], or as a core component in data auditing systems [74, 96, 107]. The issue is also highly relevant to recent ongoing lawsuits alleging unauthorized data use in model training, such as The New York Times vs. OpenAI [136], Getty Images vs. Stability AI[39] and Doe vs. GitHub[57].

Potential Mitigation Strategies Machine learning with differential privacy (DP-SGD [8] and PATE [131]) is an effective defense mechanism against privacy attacks, including MIAs. Studies [98, 135, 191] have applied DP-SGD to fine-tune LLMs on sensitive domains, which may degrade model utility under reasonable privacy budgets. For in-context learning scenarios, various DP-based techniques have been proposed, such as differential private prompt tuning [50, 70] and differential private synthetic text generation [161, 179]. In addition to the theoretical privacy guarantees that DP offers, many empirical privacy methods [71, 199] have shown effectiveness against MIAs. For example, LoRA [71], a widely used efficient fine-tuning method for LLMs, demonstrates better privacy preservation than full fine-tuning. However, these empirical defenses may be compromised when facing stronger MIAs [65], and there remains a notable trade-off between privacy and utility, especially for pre-trained LLMs.

### 3.2 Training Data Extraction

Training data extraction refers to the risk of partially or fully reconstructing samples from the training dataset by interacting with a trained LLM [22, 25, 78, 122, 146]. This threat goes beyond merely identifying whether a particular instance was part of the training set (as in membership inference); it involves recovering the training data itself, posing a more severe privacy threat. Similar to MIAs, data extraction attacks have been studied at various stages of the LLM training pipeline:

- *Pre-training Stage.* Research has shown that large amounts of data can be extracted from GPT-2 [140] by repeatedly querying it with different prompt prefixes [25]. Additionally, studies [84, 105] have demonstrated that LLMs can unintentionally leak personally identifiable information (PII), such as a person's full name, address, and phone number. More recently, studies [66] have used a probabilistic extraction approach to successfully recover pieces of copyrighted content, such as excerpts from books, from open-weight models.

- *Fine-tuned Stage.* Unlike the pre-training stage, where most data extraction attacks are conducted in a black-box setting, data extraction during fine-tuning often assumes that both the original and fine-tuned model weights are publicly available. For example, the popular reasoning model DeepSeek-R1 [43] is trained from DeepSeek-Base [42], both of which have open weights; however, the data used to train R1 has not been made public. Recent studies [116] propose efficient data selection strategies that can identify potential training data from a large data pool by matching the gradients from the base model to the fine-tuned model, demonstrating the feasibility of extracting fine-tuning data.

- *In-context Learning Stage.* Studies have shown that it is possible to recover the prompt used for in-context learning via model inversion [117] or prompt stealing [147, 160]. The privacy risks are heightened when using Retrieval-Augmented Generation (RAG) systems for LLMs, where attackers try to extract text data from a private database of RAG models through black-box access. Various attacks have been proposed to extract data from RAG systems, including adversarial prompt injection [75, 139, 170], agent-based attacks [81], and backdoor attacks [134].

Different approaches have been proposed to evaluate the effectiveness of data extraction attacks. The most widely used metric is eidetic memorization (or verbatim extraction) [25] and its variations [22, 78, 82, 164, 196]. This metric requires the model to reproduce the memorized data exactly when given an appropriate prompt. To relax this strict requirement, other studies propose approximate memorization metrics [78, 82], which measure the string or semantic similarity between the model's output and the training data as memorization efficacy.

Privacy Implications Data extraction from LLMs raises serious privacy and copyright concerns. Extraction not only produces a "copy" of training data but also reveals that a model has memorized such data internally. This evidence is central to ongoing legal debates about whether training an LLM on copyrighted material constitutes fair use. Furthermore, it poses a significant risk to the personal or proprietary data used in LLM-based applications such as RAG systems, underscoring the profound privacy challenges in deploying LLMs.

Growing Threats Recent studies [34, 66] highlight that since LLMs are probabilistic, the memorization should be assessed probabilistically. These studies introduce probabilistic discoverable extraction [66], which quantifies the likelihood that a model, under a given decoding scheme, will reproduce a verbatim target suffix when prompted with a specific prefix. This approach not only refines the understanding of memorization in LLMs but also shows how easily certain pieces of data can be extracted from these models. Further research [33] argues that measuring memorization solely by average extraction rates is insufficient. Instead, it should focus on identifying specific pieces of copyrighted or private text that are most likely to be memorized by the model. By enhancing extraction methods and pinpointing highly memorized fragments, these approaches increase the efficacy of data extraction, highlighting the potential for more targeted privacy violations.

## 4 Privacy Risks in LLM-Powered Systems

The integration of LLMs as core components in larger, complex systems introduces new vectors for privacy risks. For example, LLM-based chatbots, such as ChatGPT [127] and Claude [14], have become the primary interface through which users interact with LLMs. In these interactions, users often share personal narratives and sensitive details to seek advice or obtain personalized responses. This turns their conversation histories into a rich repository of private information, including personal preferences, habits, and even users' secrets [18, 85, 114]. Like other computer systems, these applications are vulnerable to side-channel attacks, where adversaries exploit indirect information leaks from the system to steal

data. Furthermore, the unique features of LLM-based applications, such as reasoning and memory mechanisms, provide additional attack surfaces. Below, we detail two prominent threats associated with LLM-powered systems: side channel attacks and information exfiltration.

### 4.1 Side Channel Attacks

Side-channel attacks [3, 40, 92] exploit indirect leakage of information through system behaviors such as timing, memory usage, and input/output patterns. These attacks can be especially severe in the context of LLM-based chatbots, which contain vast amounts of users' private conversation histories. Existing side-channel attacks against LLM-based chatbots can be categorized into three different types:

- *Inference Timing Attacks.* Inference timing attacks target the time it takes for an LLM to generate a response. To improve inference efficiency, modern LLMs are often optimized using data-dependent inference techniques, such as speculative decoding [112, 159], where a smaller, faster "draft" model predicts multiple future tokens, and the larger, main model verifies them in a single step. However, these optimizations introduce new vulnerabilities, as demonstrated in recent work [23, 155, 175, 193]. Specifically, the vulnerability arises from the number of drafted tokens that the main model accepts. If many tokens are accepted, the response is faster; if most are rejected, the system slows down. This acceptance rate depends on the predictability of the text. Attackers can craft specific inputs to measure these timing differences and infer the predictability of a user's hidden conversation history. By analyzing these server response time patterns, attackers can infer the topic or even specific characteristics of a user's private conversations without ever seeing the actual content.

- *Cache Timing Attacks.* Cache timing attacks exploit variations in how data is stored and accessed in a system's memory. In the context of LLM-based chatbots, inference services deployed on cloud resources need to handle a high volume of real-time requests while maintaining high throughput and low latency. One common optimization technique is prompt caching [137], where the attention key-value (KV) cache is reused across requests. In this method, the KV cache for a prompt is stored, and if a subsequent prompt shares a matching prefix with a cached prompt, the cached KV data for the prefix can be quickly retrieved. This results in faster processing times, specifically reducing the time to generate the first response token. However, the use of prompt caching introduces observable variations in response times based on the private input. When a prompt matches a cached prefix, the response is faster due to the cache hit, whereas non-matching prompts result in slower response times. By analyzing these timing differences, the attacker can learn the prefixes of other users' private inputs, potentially allowing them to identify or reconstruct the victim's entire prompt with high confidence [63, 156, 178, 208].

- *Keylogging Attacks.* Remote keylogging attacks focus on capturing the keystrokes entered by users during their interactions with chatbots. Unlike traditional keyloggers that require local access to the user's device, recent attacks [177] show it is possible to conduct this remotely by analyzing the timing and length of network packets exchanged between the user and the chatbot. By observing patterns in the size and timing of these encrypted packets, attackers can infer the length of the tokens being transmitted. Leveraging the predictable structure of language, these patterns can be used to reconstruct a user's input without any access to their device.

Privacy Implications Side-channel attacks exploit indirect signals such as inference latency, cache behavior, and packet length to infer sensitive attributes, without requiring the adversary to compromise the chatbot or the user directly. This makes them especially dangerous, as private information can be extracted passively, often without the awareness of either party.

### 4.2 Information Exfiltration

Information exfiltration refers to the unauthorized transfer of sensitive data from one context to another. In LLM-based (agent) applications, this occurs when attackers steal private information either through unintended leakage by the model itself or by maliciously manipulating the system to reveal user data, opening a new and highly exploitable attack surface [200]. We categorize existing approaches to information exfiltration into the following major categories.

- *Unintended Disclosure.* LLMs often lack awareness of privacy norms and the contextual boundaries of information flows [124]. As a result, they may inadvertently disclose sensitive information to inappropriate recipients. An LLM agent involved in a multi-round conversation may unintentionally repeat or expose previously shared user information, even when it is contextually irrelevant [38, 115, 148, 204]. For instance, it is undesirable for an LLM assistant to reveal that "John is talking to a few companies about switching jobs" while drafting an email to John's current manager, particularly without his consent. This risk increases when LLMs are tasked with complex operations that involve integrating multiple sources of user data, such as combining financial, location, and preference information for personalized recommendations [148]. LLMs struggle to track which information is appropriate to share, which makes these disclosures particularly insidious.

- *Leakage During Model Reasoning.* Recent advances in reasoning techniques encourage LLMs to generate explicit "thinking traces" or intermediate reasoning steps before producing final answers [154, 189]. While this improves task performance, studies show that reasoning traces themselves may leak sensitive user data, either accidentally or via targeted prompt injections [62]. For instance, a model assisting with medical scheduling could inadvertently include a patient's health condition in its hidden reasoning, which may later surface in outputs. This creates a difficult trade-off: increasing computational effort can make an agent's final answer more cautious, but it also encourages more verbose reasoning, thereby enlarging the attack surface.

- *Memory Leakage.* To improve personalization, many commercial LLM-powered chatbots, such as ChatGPT [127] and Gemini [162], have introduced long-term memory features that persist user information across sessions. These memories may include personal details such as location, occupation, or user preferences, stored explicitly in textual form to improve future responses. While convenient, such memory is highly sensitive and attractive to adversaries. Recent studies demonstrate that attackers can exfiltrate this data via carefully designed prompt injection attacks [133, 145]. For example, malicious content embedded in a piece of code or blog post could instruct LLM to reveal stored user memories, potentially encoding them into hidden channels (e.g., URLs or snippets of code) or misleading agents to perform actions like visiting websites that acquire user information, transferring the data to a remote adversary.

- *Insecure Tool Usage.* Tools refer to functions that LLM-based agents use to interact with external data or perform actions that modify the environment, such as writing files, clicking links on web pages, or generating and executing code. While the open-source community has made significant strides in developing secure Model Context Protocols (MCP) to ensure consistent and secure interactions with external data, these tools still pose substantial privacy risks. A recent study [35] demonstrated that MCP servers could be exploited as trojans to compromise user privacy. For example, a malicious weather MCP server, disguised as benign functionality, exploited legitimate banking tools to discover and extract user account balances. Although many vulnerabilities have been recognized [55, 88, 182], the increasing capabilities of agents with more tools at their disposal may lead to more potential privacy vulnerabilities.

- *Compromised Execution Environment.* The agent's execution environment can also be manipulated to exfiltrate sensitive information. For example, browser-based agents are highly susceptible to malicious prompt injections embedded in web pages [29, 100] or triggered by pop-ups [202], leading to the leakage of private user data. A recent study [87] further emphasizes that agents performing GUI-based tasks are particularly vulnerable, due to the misalignment between LLM behavior in conversational settings and their behavior in agent-based, browser-use contexts.

- *Leakge via Share Link.* Users may share their conversations with a chatbot (ChatGPT) through a built-in "Share" button on commercial chatbot platforms, allowing only those with the link to access the conversation. However, it has been shown [6] that these links can be discovered through search engines (Google Search). This can unintentionally expose users' conversation histories to the public. Furthermore, deleting a conversation from your ChatGPT history does not remove the public share link or prevent it from appearing in search engine results. Recognizing the potential for privacy breaches, OpenAI has addressed this issue by providing users with the option to control whether their chats are visible in search engine results, offering more control over users' privacy.

Privacy Implications Information exfiltration can occur even without the presence of an active attacker and can be unintentionally exposed through a model's internal reasoning traces or persistent memory. In this way, features originally intended to improve functionality (tool usage and share links), transparency (reasoning), and personalization (memory) become high-value attack surfaces. The ultimate consequence is a profound erosion of user trust, as interacting with a seemingly helpful LLM agent creates a persistent, exploitable record of their most sensitive information.

Vulnerability Detection and Mitigation Strategies To counter the threat of information exfiltration, research efforts are focused on both detecting vulnerabilities and developing active defenses. On the detection front, some works measure an LLM's capacity for privacy reasoning in ambiguous contexts to identify risks of unintended information disclosure [149, 190]. Another approach [16, 206] examines whether agents interacting with web interfaces adhere to the principle of data minimization, introducing benchmarks to systematically evaluate their compliance. In parallel, other efforts aim to build direct defenses against malicious attacks. This includes designing robust countermeasures for prompt injection attacks, which are a primary vector for information exfiltration [41, 168]. Despite these advancements, a comprehensive mitigation strategy capable of defending against the full spectrum of emerging exfiltration threats remains an open challenge [67, 200].

## 5 Privacy Risks from Malicious Use of LLMs

The increasingly impressive capabilities of LLMs have demonstrated remarkable potential across diverse fields, such as software engineering [187], human behavior simulation [132], and even assisting scientific discovery [17]. However, this progress presents a dual-use dilemma, as the very capabilities driving these innovations can also be misused for malicious purposes [24, 28, 56, 91, 183]. Specifically, LLMs amplify the risk of privacy violations in two ways:

- *Scaling Sophisticated Attacks.* LLMs can automate and execute privacy attacks that were previously prohibitive due to their complexity or high cost. By either assisting human adversaries or operating independently, they can enable privacy breaches at an unprecedented scale.

- *Democratizing Attack Capabilities.* LLMs lower the barrier for malicious actors by making powerful attack tools accessible to people with little to no expertise. This "democratization" allows individuals with limited knowledge to launch attacks that previously required specialized skills.

In this section, we introduce two emerging privacy risks of the malicious use of LLMs: automated profile inference and automated social engineering.

### 5.1 Automated Profile Inference

Individuals constantly generate digital footprints through their online activities, encompassing activities from social media comments and posts to shared photos and videos. While some of this data is inherently private (browse history), a vast amount of these activities (posts and comments) is publicly accessible. However, the public availability of this information does not eliminate privacy risks. An adversary can aggregate these seemingly innocuous public activities to construct a detailed personal profile, a process known as profiling [20, 47]. For instance, analyzing a Reddit user's most frequented subreddits could reveal their hobbies, while geotags in posted images could disclose their travel patterns or home location. Profiling is widely recognized as a privacy violation by legitimate privacy frameworks like GDPR[142], CCPA [126], and HIPAA [9].

Profiling based on unstructured and noisy data requires significant expertise and is considered too resource-intensive for large-scale privacy breaches [46]. The emergence of LLMs fundamentally alters this landscape. By leveraging their sophisticated understanding and reasoning capabilities, LLMs can automate the inference process, systematically analyzing vast digital footprints to infer sensitive attributes with minimal human intervention. This automation dramatically amplifies the threat, enabling profiling attacks at an unprecedented scale. A growing body of work has demonstrated the feasibility of LLM-driven profiling attacks [49, 158, 165]. In the following, we categorize these attacks along two primary axes: (i) the data modality they target and (ii) their level of automation.

Profiling Across Data Modalities With the increasing capabilities of LLMs in understanding different data modalities, various profiling attacks have been proposed by analyzing a user's activities across multiple types of data:

- *Profiling from Textual Activities.* Early LLM-based profiling attack [158] assumes that an adversary can access and scrape the public activities (posts and comments) of a pseudonymous user from the Internet. The adversary then instructs LLMs with prompts to infer predefined sensitive attributes (eight types of PIIs), within these textual activities. The results showed that powerful models like GPT-4 [127] can achieve performance comparable to human analysts, even when the humans have the advantage of accessing additional contextual information, which LLMs do not have.

- *Profiling from Visual Activities.* With the rise of Vision-Language Models (VLMs) [128, 162], research has expanded to include profiling from images and videos, which are ubiquitous on social media platforms like TikTok and Instagram. Specifically, one study [165] designed carefully crafted prompts using chain-of-thought reasoning [174] and automated zooming techniques to direct VLMs to focus on potentially sensitive details in the photos, thus enhancing privacy-infringing inferences. Another significant privacy risk arises from directly inferring a user's possible geo-location from their pictures [73, 80, 102, 188]. Research has shown that VLMs can outperform even the professional human players in GeoGuessr [2], which raises serious concerns regarding geographic privacy. However, these models are not infallible; they often exhibit significant regional biases, such as a tendency to over-predict well-known landmarks or locations heavily represented in their training data [73].

Different Levels of Automation in Profiling The privacy risks associated with the malicious use of LLMs depend heavily on the degree of automation involved in the attack. A highly automated and practical attack poses a much greater real-world privacy threat, as it reduces the need for human

adversaries, making it more cost-efficient and scalable. We categorize existing approaches into two types: semi-automated and fully automated, depending on their level of automation:

- *Semi-Automated Profiling.* The majority of current research falls into this category, where the core inference task is automated, but significant human effort is still required for data preparation and defining attack objectives. These systems are powerful in controlled settings but face two major limitations in real-world scenarios: (i) Reliance on curated data. Many studies [104, 158, 165, 188] focus on clean, curated textual or image data that is deliberately designed to contain sensitive information, allowing LLMs and VLMs to infer personal attributes. However, in real-world scenarios, user activities are typically noisy and may not be directly related to personal attributes. As a result, the performance of these semi-automated methods would likely degrade significantly when faced with raw, unfiltered activities. (ii) Predefined attribute targets. These attacks are typically configured to search for a fixed set of sensitive attributes (age, gender, location), which assumes the adversary already knows what to profile from users. However, in the real world, adversaries do not always know what sensitive attributes are present in a user's activities. This lack of predefined knowledge prevents the attacker from designing specific strategies to target particular attributes, further limiting the applicability of such attacks.

- *Fully-Automated Profiling.* To address the limitations of previous approaches, recent work has focused on developing end-to-end automated profiling systems. One example is AutoProfiler [49], an agent-based profiling framework that automatically scrapes, collects, and analyzes potentially sensitive activities from raw, noisy user data. By coordinating with four specialized LLM agents, AutoProfiler fully automates the process of inferring sensitive attributes. This eliminates the need for background knowledge or profiling expertise, making it highly scalable and suitable for deployment on web-scale platforms. Despite its weaker assumptions, the results show that the inferred attributes extend beyond PIIs, uncovering significant amounts of sensitive information. The move toward full automation has profound implications. It means that adversaries no longer need specialized expertise or prior knowledge to launch sophisticated, large-scale profiling attacks.

Privacy Implications Automated profiling inference can result in serious privacy breaches. One of the most well-known risks is de-anonymization [120, 121]. Study [49] shows that some Reddit users can be de-anonymized by inferring personal attributes from their public activities and comparing these with publicly available profiles, such as LinkedIn. The risk of de-anonymization increases when adversaries gain access to multiple profile databases or cross-reference a user's activities to construct more comprehensive profiles. In addition, sensitive information extracted from these online activities can also be exploited for severe cybercrimes like doxing and cyberbullying. We refer to [46, 49] for a deeper discussion of the consequences of exposing sensitive personal data.

Growing Threats Existing attacks exploit the in-context learning (ICL) capability of off-the-shelf LLMs to perform profiling tasks. While this approach is highly efficient and accessible, its performance could be suboptimal, as these models are not specifically designed for profiling. For example, studies show that even state-of-the-art VLMs are outperformed in geo-location identification tasks by PIGEON [64], an image model purpose-built for geolocation. This trend suggests that adversaries may design specialized profiling models that surpass generic LLMs, thereby enhancing attack effectiveness and posing even more severe privacy risks.

Challenges in Evaluation While various methodologies have been proposed to assess the profiling abilities of LLMs [49, 73, 158, 183], there is still no widely acknowledged benchmark to comprehensively evaluate the associated privacy risks. This issue partially stems from a fundamental ethical dilemma: creating a robust benchmark would require a large dataset of real users' activities with labeled, sensitive, ground-truth attributes. To address this, some researchers have proposed using synthetic datasets

generated by LLM agents [79, 192]. However, the behaviors and data produced by these agents may not accurately reflect the complexities of real human activity, limiting their validity and reliability [49]. In addition, evaluation becomes more complex for fully automated systems that perform open-ended inference without predefined attribute targets. Forcing the model to choose from a candidate list simplifies evaluation but fails to measure the model's true, unconstrained inference capabilities. Therefore, evaluating the profiling abilities of LLMs remains an open question. Designing effective evaluation approaches is a critical step toward understanding and mitigating these emerging privacy threats.

### 5.2 Automated Social Engineering

A social engineering attack exploits the psychological manipulation of human behavior to extract sensitive information, gain access to personal devices, share credentials, or perform other malicious activities that compromise digital security [125]. There are different types of social engineering, such as phishing, vishing, pretexting, and baiting. Over the past decades, social engineering attacks have resulted in numerous incidents, causing severe financial losses and privacy breaches [60, 123]. Most social engineering attacks follow four main stages [141]: (i) *Investigation.* The attacker gathers information about the target, often from public social media, job platforms, and online sources, to identify vulnerabilities. (ii) *Planning.* Based on the gathered information, the attacker develops a strategy, selecting tactics like phishing or impersonation to exploit weaknesses. (iii) *Contact.* The attacker establishes trust with the target, persuading them to take harmful actions such as clicking a malicious link or disclosing sensitive information. (iv) *Execution.* The attacker extracts sensitive data, installs malware, or otherwise compromises the target's system.

Social engineering attacks typically required significant human effort and expertise, and their success rates were often limited by defense mechanisms and human vigilance. For example, phishing emails could be easily recognized by telltale signs like grammatical errors or implausible scenarios [89]. However, the advent of LLMs has introduced a new dimension to social engineering threats, which we refer to as automated social engineering. Unlike traditional methods, LLM-driven attacks can be personalized and executed at scale. These models can automate and enhance all four major stages of a social engineering attack, increasing both effectiveness and efficiency, as detailed below.

Automated Investigation The purpose of this phase is to gather sufficient information about a target to personalize the attack and make it more convincing [123]. Adversaries may directly employ automated profiling strategies (as described in the previous section) to collect personal information. In addition, they may launch proactive information-gathering attempts by manipulating LLM-based chatbots to elicit sensitive details. In such scenarios, a chatbot convinces the user that certain personal information is required to complete a task. Because users often perceive LLMs as helpful assistants, they may willingly provide sensitive details, believing them to be necessary [12, 86, 90, 195]. Attackers can exploit this trust by embedding hidden, privacy-invasive prompts into a chatbot's behavior [158]. For example, a chatbot tasked with creating a travel itinerary might subtly request additional personal details—such as financial information or contact numbers—under the guise of improving the service. The risk is further amplified in multi-agent LLM systems, where multiple agents collaborate by asking for complementary pieces of information and together constructing a detailed personal profile of the victim [209]. These LLM-based information collection strategies dramatically reduce the cost and time required for reconnaissance while producing highly detailed and actionable intelligence about targets.

LLM-Aided Planning In this stage, LLMs could serve as powerful reasoning and analysis engines to help attackers design persuasive attack strategies. Specifically, LLMs can (i) propose tailored attack vectors—such as spear-phishing campaigns or impersonation scenarios, (ii) generate dialogue templates to sustain orchestrated interactions that gradually build trust [163], and (iii) dynamically adapt strategies, for example by suggesting follow-up messages when a target hesitates or fails to respond.

This capability transforms attack planning from a manual, experience-driven art into an automated process. Sophisticated, customized attack blueprints can be generated in minutes, removing the need for an experienced human attacker.

LLM-Enhanced Contact LLMs can be exploited not only to enhance interactions through existing contact channels but also to create entirely new avenues for reaching targets.

First, LLMs enhance traditional methods like phishing by generating persuasive, context-aware emails with remarkable speed. Studies show an LLM could draft a highly convincing spear-phishing email in just five minutes, a task that takes a human team several hours [31, 76, 77]. LLMs can also sustain convincing, real-time conversations that gradually build trust. When paired with generative deepfake technologies for images, video, or audio, impersonations become nearly indistinguishable from legitimate contacts [10, 89, 167]. This allows a single attacker to maintain persistent, personalized engagement across multiple platforms and scale their outreach to thousands of potential victims simultaneously.

Second, LLMs open new avenues for attack by exploiting the growing use of chatbots for emotional and psychological support [72, 207]. In this scenario, attackers deploy malicious chatbots that impersonate trusted friends or companions to establish a deep emotional connection with a victim. The proliferation of third-party platforms like the OpenAI GPT Store [4] and FlowGPT [5] makes it easy to distribute these deceptive chatbots to a wide audience. Once an emotional connection is established, adversaries can manipulate victims into disclosing sensitive information, transferring money under fraudulent pretenses, or even engaging in harmful behaviors [7].

LLM-Aided Execution Once trust is established and sensitive data is obtained, LLMs can assist attackers in carrying out malicious actions. These include: (i) leveraging stolen credentials to gain unauthorized access to systems [184], (ii) automating financial fraud, such as wire transfer scams [106], and (iii) orchestrating follow-on attacks, including malware distribution or pivoting to additional targets within a compromised network [143]. By reducing the need for manual effort, LLMs enable end-to-end, scalable, and highly sophisticated attack pipelines.

Privacy and Security Implications Automated social engineering represents a multifaceted threat to both privacy and security. It dramatically increases the risk of large-scale data leakage and financial loss. Attackers can harvest sensitive personal information, financial details, and corporate credentials with unprecedented efficiency [76]. The real-world consequences are staggering; in one recent incident, fraudsters used a combination of phishing and video-based deepfake impersonation to deceive an employee into authorizing a fraudulent $25 million transfer [106]. Beyond financial loss, certain strategies exploit users' trust or emotional reliance, inflicting psychological harm that can result in profound emotional distress. Thus, automated social engineering not only increases the efficiency of attacks but also expands the pool of potential victims, thereby amplifying the societal impact of privacy breaches.

Growing Threats With the rapid development of LLMs, automated social engineering attacks may become even more sophisticated and hard to detect. Multi-modal LLMs, for example, can generate coordinated text, audio, and video content to produce highly immersive impersonations that are nearly indistinguishable from authentic human interactions. Another concern lies in the emergence of autonomous agents capable of orchestrating end-to-end attack campaigns. Such agents could handle reconnaissance, planning, multi-turn conversations, and final exploitation without any human oversight [89]. These advancements suggest that future LLM-driven social engineering would progress beyond opportunistic scams toward coordinated, persistent, and large-scale operations capable of evading even advanced detection and defense systems.

Vulnerability Detection and Mitigation Strategies Several studies have examined the capabilities of LLMs in conducting social engineering and their impact on human users [69, 101, 184]. For example, a recent work [130] proposed embedding trigger–tag associations into vanilla LLMs through various insertion strategies. When the model is instructed to generate phishing emails, detectable tags are inserted into the output, enabling more effective detection of LLM-generated phishing content. However,

such safety enhancements for LLMs are limited in their real-world applicability. Adversaries can easily bypass them by locally deploying open-source and unconstrained LLMs without these safeguards. Thus, the challenge extends beyond detecting LLM-generated social engineering content to also identifying autonomous malicious activities carried out by LLM agents.

## 6 Conclusion

The rapid development and integration of LLMs into digital infrastructure and daily life have introduced a new frontier of privacy risks. In this paper, we systematically examined emerging threats of LLMs across three dimensions: (i) data privacy risks across various learning stages of LLMs; (ii) privacy risks in LLM-powered applications, including side channels and information exfiltration; and (iii) malicious use of LLMs, such as automated profiling and social engineering. We then discuss the real-world privacy implications of these threats and highlight the limitations of existing mitigation strategies. This paper helps to illuminate the privacy risks introduced by LLMs and advocates for greater social awareness of these challenges. We also call for research efforts that broaden their focus beyond data privacy and design new defenses to address these privacy threats.

## References


[1] Chatgpt. https://chatgpt.com/.

[2] GeoGuessr. https://www.geoguessr.com/.

[3] Side-channel attack. https://en.wikipedia.org/wiki/Side-channel_attack.

[4] GPT Store. https://gptstore.ai/, 2023.

[5] FlowGPT. https://flowgpt.com/, 2025.

[6] OpenAI Is Pulling Shared ChatGPT Chats From Google Search. https://www.searchenginejournal.com/openai-is-pulling-shared-chatgpt-chats-from-google-search 552671/, 2025.

[7] Parents of teenager who took his own life sue OpenAI. https://www.bbc.com/news/articles/cgerwp7rdlvo, 2025.

[8] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[9] Accountability Act. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996.

[10] Lin Ai, Tharindu Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, et al. Defending against social engineering attacks in the age of llms. *arXiv preprint arXiv:2406.12263*, 2024.

[11] NIST AI. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, 2024.

[12] Mutahar Ali, Arjun Arunasalam, and Habiba Farrukh. Understanding Users' Security and Privacy Concerns and Attitudes Towards Conversational AI Platforms. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 298–316, 2025.

[13] Maya Anderson, Guy Amit, and Abigail Goldsteen. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*, 2024.

[14] Anthropic. Claude.ai. `https://claude.ai`.

[15] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. 2024.

[16] Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. Airgapagent: Protecting privacy-conscious conversational agents. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3868–3882, 2024.

[17] Daniil A Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.

[18] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 2280–2292, 2022.

[19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[20] Firearms Bureau of Alcohol, Tobacco and Explosives. Criminal Profilers.

[21] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022.

[22] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[23] Nicholas Carlini and Milad Nasr. Remote timing attacks on efficient language model inference. *arXiv preprint arXiv:2410.17175*, 2024.

[24] Nicholas Carlini, Milad Nasr, Edoardo Debenedetti, Barry Wang, Christopher A Choquette-Choo, Daphne Ippolito, Florian Tramèr, and Matthew Jagielski. LLMs unlock new paths to monetizing exploits. *arXiv preprint arXiv:2505.11449*, 2025.

[25] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[26] 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 *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[27] Hongyan Chang, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Reza Shokri. Context-aware membership inference attacks against pre-trained large language models. *arXiv preprint arXiv:2409.13745*, 2024.

[28] Yaqub Chaudhary and Jonnie Penn. Large language models as instruments of power: New regimes of autonomous manipulation and control. *arXiv preprint arXiv:2405.03813*, 2024.

[29] Chaoran Chen, Zhiping Zhang, Bingcan Guo, Shang Ma, Ibrahim Khalilov, Simret A Gebreegziabher, Yanfang Ye, Ziang Xiao, Yaxing Yao, Tianshi Li, et al. The Obvious Invisible Threat: LLM-Powered GUI Agents' Vulnerability to Fine-Print Injections. *arXiv preprint arXiv:2504.11281*, 2025.

[30] Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-jun Li, and Yaxing Yao. Clear: Towards contextual llm-empowered privacy policy analysis and risk generation for large language model applications. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 277–297, 2025.

[31] Fengchao Chen, Tingmin Wu, Van Nguyen, Shuo Wang, Hongsheng Hu, Alsharif Abuadbba, and Carsten Rudolph. Adapting to Cyber Threats: A Phishing Evolution Network (PEN) Framework for Phishing Generation and Analyzing Evolution Patterns using Large Language Models. *arXiv preprint arXiv:2411.11389*, 2024.

[32] 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 *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1285–1299, 2024.

[33] A Feder Cooper, Aaron Gokaslan, Amy B Cyphert, Christopher De Sa, Mark A Lemley, Daniel E Ho, and Percy Liang. Extracting memorized pieces of (copyrighted) books from open-weight language models. *arXiv preprint arXiv:2505.12546*, 2025.

[34] A. Feder Cooper and James Grimmelmann. The Files are in the Computer: Copyright, Memorization, and Generative AI. *arXiv preprint arXiv:2404.12590*, 2024.

[35] Nicola Croce and Tobin South. Trivial Trojans: How Minimal MCP Servers Enable Cross-Tool Exfiltration of Sensitive Data. *arXiv preprint arXiv:2507.19880*, 2025.

[36] Cursor. Cursor Team. `https://www.cursor.com/`.

[37] Debeshee Das, Jie Zhang, and Florian Trantèr. Blind Baselines Beat Membership Inference Attacks for Foundation Models. In *2025 IEEE Security and Privacy Workshops*, pages 118–125, 2025.

[38] Saswat Das, Jameson Sandler, and Ferdinando Fioretto. Disclosure Audits for LLM Agents. *arXiv preprint arXiv:2506.10171*, 2025.

[39] Cerys Wyn Davies and Gill Dennis. Getty Images v Stability AI: the implications for UK copyright law and licensing. `https://www.pinsentmasons.com/out-law/analysis/getty-images-v-stability-ai-implications-copyright-law-licensing`.

[40] Edoardo Debenedetti, Giorgio Severi, Nicholas Carlini, Christopher A Choquette-Choo, Matthew Jagielski, Milad Nasr, Eric Wallace, and Florian Tramèr. Privacy side channels in machine learning systems. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 6861–6848, 2024.

[41] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design. *arXiv preprint arXiv:2503.18813*, 2025.

[42] DeepSeek-AI. DeepSeek-V3 Technical Report. 2024.

[43] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. 2025.

[44] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

[45] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[46] David M Douglas. Doxing: a conceptual analysis. *Ethics and information technology*, 18(3):199–210, 2016.

[47] John E Douglas, Robert K Ressler, Ann W Burgess, and Carol R Hartman. Criminal profiling from crime scene analysis. *Behavioral Sciences & the Law*, 4(4):401–421, 1986.

[48] Elias Dritsas, Maria Trigka, and Phivos Mylonas. A Survey on Privacy-Enhancing Techniques in the Era of Artificial Intelligence. In *Novel & Intelligent Digital Systems Conferences*, pages 385–392, 2024.

[49] Yuntao Du, Zitao Li, Bolin Ding, Yaliang Li, Hanshen Xiao, Jingren Zhou, and Ninghui Li. Automated Profile Inference with Language Model Agents. In *Workshop on AI Agents: Capabilities and Safety, Conference on Language Modeling (COLM)*, 2025.

[50] Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. *Advances in Neural Information Processing Systems*, 36:76852–76871, 2023.

[51] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do Membership Inference Attacks Work on Large Language Models? In *Conference on Language Modeling (COLM)*, 2024.

[52] André Vicente Duarte, Xuandong Zhao, Arlindo L Oliveira, and Lei Li. DE-COP: Detecting Copyrighted Content in Language Models Training Data. In *International Conference on Machine Learning*, pages 11940–11956, 2024.

[53] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[54] Cynthia Dwork. Differential Privacy. In *Automata, Languages and Programming*, pages 1–12, 2006.

[55] Junfeng Fang, Zijun Yao, Ruipeng Wang, Haokai Ma, Xiang Wang, and Tat-Seng Chua. We Should Identify and Mitigate Third-Party Safety Risks in MCP-Powered Agent Systems. *arXiv preprint arXiv:2506.13666*, 2025.

[56] Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. LLM agents can autonomously hack websites. *arXiv preprint arXiv:2402.06664*, 2024.

[57] Jose Florinio Farcon. Attribution Or Attrition? Doe 1 V. Github, Inc. As A Case For A Robust, Horizontal, Moral Right Of Attribution In Gen AI. 2024.

[58] Qizhang Feng, Siva Rajesh Kasa, SANTHOSH KUMAR KASA, Hyokun Yun, Choon Hui Teo, and Sravan Babu Bodapati. Exposing Privacy Gaps: Membership Inference Attack on Preference Data for LLM Alignment. In *International Conference on Artificial Intelligence and Statistics*, pages 5221–5229. PMLR, 2025.

[59] James Flemings, Bo Jiang, Wanrong Zhang, Zafar Takhirov, and Murali Annavaram. Estimating Privacy Leakage of Augmented Contextual Knowledge in Language Models. *arXiv preprint arXiv:2410.03026*, 2025.

[60] World Economic Forum. AI could empower and proliferate social engineering cyberattacks, 2024.

[61] Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership Inference Attacks against Fine-tuned Large Language Models via Self-prompt Calibration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[62] Tommaso Green, Martin Gubri, Haritz Puerto, Sangdoo Yun, and Seong Joon Oh. Leaky Thoughts: Large Reasoning Models Are Not Private Thinkers. *arXiv preprint arXiv:2506.15674*, 2025.

[63] Chenchen Gu, Xiang Lisa Li, Rohith Kuditipudi, Percy Liang, and Tatsunori Hashimoto. Auditing Prompt Caching in Language Model APIs. *arXiv preprint arXiv:2502.07776*, 2025.

[64] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12893–12902, 2024.

[65] Jamie Hayes, Ilia Shumailov, Christopher A Choquette-Choo, Matthew Jagielski, George Kaissis, Katherine Lee, Milad Nasr, Sahra Ghalebikesabi, Niloofar Mireshghallah, Meenatchi Sundaram Mutu Selva Annamalai, et al. Strong Membership Inference Attacks on Massive Datasets and (Moderately) Large Language Models. *arXiv preprint arXiv:2505.18773*, 2025.

[66] Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, Ilia Shumailov, Milad Nasr, Christopher A. Choquette-Choo, Katherine Lee, and A. Feder Cooper. Measuring memorization in language models via probabilistic extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 9266–9291, 2025.

[67] Pengfei He, Yue Xing, Shen Dong, Juanhui Li, Zhenwei Dai, Xianfeng Tang, Hui Liu, Han Xu, Zhen Xiang, and Charu C Aggarwal. Comprehensive Vulnerability Analysis is Necessary for Trustworthy LLM-MAS. *arXiv preprint arXiv:2506.01245*, 2025.

[68] Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, Yiming Li, Zhan Qin, Kui Ren, and Chun Chen. Towards label-only membership inference attack against pre-trained large language models. In *USENIX Security*, 2025.

[69] Fred Heiding, Simon Lermen, Andrew Kao, Bruce Schneier, and Arun Vishwanath. Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

[70] Junyuan Hong, Jiachen T Wang, Chenhui Zhang, Zhangheng LI, Bo Li, and Zhangyang Wang. Dp-opt: Make large language model your privacy-preserving prompt engineer. In *International Conference on Learning Representations*, 2024.

[71] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[72] Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, 8(1):230, 2025.

[73] Jingyuan Huang, Jen-tse Huang, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang, and Jieyu Zhao. Vlms as geoguessr masters: Exceptional performance, hidden biases, and privacy risks. *arXiv preprint arXiv:2502.11163*, 2025.

[74] Zonghao Huang, Neil Zhenqiang Gong, and Michael K Reiter. A general framework for data-use auditing of ML models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1300–1314, 2024.

[75] Bo Hui, Haolin Yuan, Neil Gong, Philippe Burlina, and Yinzhi Cao. PLeak: Prompt Leaking Attacks against Large Language Model Applications. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3600–3614, 2024.

[76] IBM. AI vs. human deceit: Unravelling the new age of phishing tactics. `https://www.ibm.com/think/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics`, 2023.

[77] IBM. With generative AI, social engineering gets more dangerous—and harder to spot. `https://www.ibm.com/think/insights/generative-ai-social-engineering`, 2025.

[78] Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, 2023.

[79] Shalini Jangra, Suparna De, Nishanth Sastry, and Saeed Fadaei. Protecting Vulnerable Voices: Synthetic Dataset Generation for Self-Disclosure Detection. *arXiv preprint arXiv:2507.22930*, 2025.

[80] Neel Jay, Hieu Minh Nguyen, Trung-Dung Hoang, and Jacob Haimes. Evaluating Precise Geolocation Inference Capabilities of Vision Language Models. 2025.

[81] Changyue Jiang, Xudong Pan, Geng Hong, Chenfu Bao, Yang Chen, and Min Yang. Feedback-Guided Extraction of Knowledge Base from Retrieval-Augmented LLM Applications. 2025.

[82] 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, 2022.

[83] Gyuwan Kim, Yang Li, Evangelia Spiliopoulou, Jie Ma, Miguel Ballesteros, and William Yang Wang. Detecting training data of large language models via expectation maximization. *arXiv preprint arXiv:2410.07582*, 2024.

[84] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. ProPILE: probing privacy leakage in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

[85] Isadora Krsek, Anubha Kabra, Yao Dou, Tarek Naous, Laura A. Dabbish, Alan Ritter, Wei Xu, and Sauvik Das. Measuring, Modeling, and Helping People Account for Privacy Risks in Online Self-Disclosures with AI. *Proc. ACM Hum. Comput. Interact.*, 9(2):1–31, 2025.

[86] Isadora Krsek, Anubha Kabra, Yao Dou, Tarek Naous, Laura A. Dabbish, Alan Ritter, Wei Xu, and Sauvik Das. Measuring, Modeling, and Helping People Account for Privacy Risks in Online Self-Disclosures with AI. *Proc. ACM Hum.-Comput. Interact.*, 2025.

[87] Priyanshu Kumar, Elaine Lau, Saranya Vijayakumar, Tu Trinh, Elaine T Chang, Vaughn Robinson, Shuyan Zhou, Matt Fredrikson, Sean M Hendryx, Summer Yue, et al. Aligned LLMs are not aligned browser agents. In *The Thirteenth International Conference on Learning Representations*, 2025.

[88] Sonu Kumar, Anubhav Girdhar, Ritesh Patil, and Divyansh Tripathi. Mcp guardian: A security-first layer for safeguarding mcp-based ai system. *arXiv preprint arXiv:2504.12757*, 2025.

[89] Tharindu Kumarage, Cameron Johnson, Jadie Adams, Lin Ai, Matthias Kirchner, Anthony Hoogs, Joshua Garland, Julia Hirschberg, Arslan Basharat, and Huan Liu. Personalized Attacks of Social Engineering in Multi-turn Conversations–LLM Agents for Simulation and Detection. *arXiv preprint arXiv:2503.15552*, 2025.

[90] Jabari Kwesi, Jiaxun Cao, Riya Manchanda, and Pardis Emami-Naeini. Exploring User Security and Privacy Attitudes and Concerns Toward the Use of {General-Purpose}{LLM} Chatbots for Mental Health. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 6007–6024, 2025.

[91] Marcus Law. Scam email cyber attacks increase after rise of ChatGPT. *Technology*, 2023.

[92] Liran Lerman, Gianluca Bontempi, and Olivier Markowitch. Side channel attack: an approach based on machine learning. *Center for Advanced Security Research Darmstadt*, 29:29–41, 2011.

[93] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[94] Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. Privacy in Large Language Models: Attacks, Defenses and Future Directions. 2023.

[95] Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu, and Weining Yang. Membership privacy: A unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 889–900, 2013.

[96] 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. *Proceedings of the VLDB Endowment*, 17(11):3201–3214, 2024.

[97] Tianshi Li, Sauvik Das, Hao-Ping (Hank) Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. Human-Centered Privacy Research in the Age of Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 581:1–581:4, 2024.

[98] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large Language Models Can Be Strong Differentially Private Learners. In *International Conference on Learning Representations*, 2022.

[99] Yiming Li, Shuo Shao, Yu He, Junfeng Guo, Tianwei Zhang, Zhan Qin, Pin-Yu Chen, Michael Backes, Philip Torr, Dacheng Tao, et al. Rethinking data protection in the (generative) artificial intelligence era. *arXiv preprint arXiv:2507.03034*, 2025.

[100] Zeyi Liao, Lingbo Mo, Chejian Xu, Mintong Kang, Jiawei Zhang, Chaowei Xiao, Yuan Tian, Bo Li, and Huan Sun. EIA: ENVIRONMENTAL INJECTION ATTACK ON GENERALIST WEB AGENTS FOR PRIVACY LEAKAGE. In *The Thirteenth International Conference on Learning Representations*.

[101] Zilong Lin, Jian Cui, Xiaojing Liao, and XiaoFeng Wang. Malla: Demystifying real-world large language model integrated malicious services. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4693–4710, 2024.

[102] Feiran Liu, Yuzhe Zhang, Xinyi Huang, Yinan Peng, Xinfeng Li, Lixu Wang, Yutong Shen, Ranjie Duan, Simeng Qin, Xiaojun Jia, Qingsong Wen, and Wei Dong. The Eye of Sherlock Holmes: Uncovering User Private Attribute Profiling via Vision-Language Model Agentic Framework. abs/2505.19139, 2025.

[103] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024.

[104] Yupei Liu, Yuqi Jia, Jinyuan Jia, and Neil Zhenqiang Gong. Evaluating LLM-based Personal Information Extraction and Countermeasures. In *34st USENIX Security Symposium (USENIX Security 25)*, 2025.

[105] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing Leakage of Personally Identifiable Information in Language Models. In *44th IEEE Symposium on Security and Privacy*, pages 346–363, 2023.

[106] Kathleen Magramo. British engineering giant Arup revealed as 25 million deepfake scam victim, 2024.

[107] Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLM Dataset Inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092, 2024.

[108] Nestor Maslej, Loredana Fattorini, C. Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. Artificial Intelligence Index Report 2025. *arXiv preprint arXiv:2504.07139*, 2025.

[109] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership Inference Attacks against Language Models via Neighbourhood Comparison. In *Findings of the Association for Computational Linguistics*, pages 11330–11343, 2023.

[110] Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2369–2385, 2024.

[111] Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. Sok: Membership inference attacks on llms are rushing nowhere (and how to fix it). In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 385–401, 2025.

[112] Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, et al. Specinfer: Accelerating large language model serving with tree-based speculative inference and verification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 932–949, 2024.

[113] Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David K Evans, and Taylor Berg-Kirkpatrick. 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, 2022.

[114] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *The Twelfth International Conference on Learning Representations*.

[115] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *The Twelfth International Conference on Learning Representations*, 2024.

[116] John X Morris, Junjie Oscar Yin, Woojeong Kim, Vitaly Shmatikov, and Alexander M Rush. Approximating Language Model Training Data from Weights. *arXiv preprint arXiv:2506.15553*, 2025.

[117] John Xavier Morris, Wenting Zhao, Justin T Chiu, Vitaly Shmatikov, and Alexander M Rush. Language Model Inversion. In *The Twelfth International Conference on Learning Representations*, 2024.

[118] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

[119] Sasi Kumar Murakonda and Reza Shokri. ML privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. *arXiv preprint arXiv:2007.09339*, 2020.

[120] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *29th IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

[121] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *30th IEEE symposium on security and privacy*, pages 173–187, 2009.

[122] Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable Extraction of Training Data from Aligned, Production Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[123] Anam Naz, Muhammad Sarwar, Muhammad Kaleem, Muhammad Azhar Mushtaq, and Salman Rashid. A comprehensive survey on social engineering-based attacks on social networks. *International Journal of Advanced and Applied Sciences*, 11(4):139–154, 2024.

[124] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

[125] Information Technology Laboratory (NIST). Social Engineering, 2024.

[126] State of California Legislature. California Consumer Privacy Act of 2018. *Public law*, 2018.

[127] OpenAI. GPT-4 Technical Report, 2023.

[128] OpenAI. GPT-4V(ision) System Card, 2023.

[129] Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[130] Yan Pang, Wenlong Meng, Xiaojing Liao, and Tianhao Wang. Paladin: Defending LLM-enabled Phishing Emails with a New Trigger-Tag Paradigm. *arXiv preprint arXiv:2509.07287*, 2025.

[131] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2017.

[132] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[133] Atharv Singh Patlan, S Ashwin Hebbar, Pramod Viswanath, and Prateek Mittal. Context manipulation attacks: Web agents are susceptible to corrupted memory. In *ICML 2025 Workshop on Computer Use Agents*.

[134] Yuefeng Peng, Junda Wang, Hong Yu, and Amir Houmansadr. Data extraction attacks in retrieval-augmented generation via backdoors. *arXiv preprint arXiv:2411.01705*, 2024.

[135] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023.

[136] Audrey Pope. NYT v. OpenAI: The Times's About-Face. `https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/`.

[137] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624, 2023.

[138] Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. Scaling up membership inference: When and how attacks succeed on large language models. *arXiv preprint arXiv:2411.00154*, 2024.

[139] Zhenting Qi, Hanlin Zhang, Eric P. Xing, Sham M. Kakade, and Himabindu Lakkaraju. Follow My Instruction and Spill the Beans: Scalable Data Extraction from Retrieval-Augmented Generation Systems. In *The Thirteenth International Conference on Learning Representations*, 2025.

[140] 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.

[141] Tejal Rathod, Nilesh Kumar Jadav, Sudeep Tanwar, Abdulatif Alabdulatif, Deepak Garg, and Anupam Singh. A comprehensive survey on social engineering attacks, countermeasures, case study, and research challenges. *Information Processing & Management*, 62(1):103928, 2025.

[142] Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (EU)*, 679, 2016.

[143] MIT Technology Review. Cyberattacks by AI agents are coming. `https://www.technologyreview.com/2025/04/04/1114228/cyberattacks-by-ai-agents-are-coming/`, 2025.

[144] Erik Schluntz and Barry Zhang. Building effective agents. `https://www.anthropic.com/research/building-effective-agents`, 2024. Anthropic.

[145] Gregory Schwartzman. Exfiltration of personal information from ChatGPT via prompt injection. *arXiv preprint arXiv:2406.00199*, 2024.

[146] Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary Lipton, and J Zico Kolter. Rethinking llm memorization through the lens of adversarial compression. *Advances in Neural Information Processing Systems*, 37:56244–56267, 2024.

[147] Zeyang Sha and Yang Zhang. Prompt Stealing Attacks Against Large Language Models. 2024.

[148] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[149] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. Privacylens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems*, 37:89373–89407, 2024.

[150] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[151] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

[152] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18, 2017.

[153] Victoria Smith, Ali Shahin Shamsabadi, Carolyn Ashurst, and Adrian Weller. Identifying and Mitigating Privacy Risks Stemming from Language Models: A Survey. 2023.

[154] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[155] Mahdi Soleimani, Grace Jia, In Gim, Seung-seob Lee, and Anurag Khandelwal. Wiretapping LLMs: Network side-channel attacks on interactive LLM services. *Cryptology ePrint Archive*, 2025.

[156] Linke Song, Zixuan Pang, Wenhao Wang, Zihao Wang, XiaoFeng Wang, Hongbo Chen, Wei Song, Yier Jin, Dan Meng, and Rui Hou. The early bird catches the leak: Unveiling timing side channels in llm serving systems. *arXiv preprint arXiv:2409.20002*, 2024.

[157] Shuang Song and David Marn. Introducing a New Privacy Testing Library in TensorFlow. 2022.

[158] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024.

[159] Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31, 2018.

[160] Yicong Tan, Xinyue Shen, Yun Shen, Michael Backes, and Yang Zhang. On the Effectiveness of Prompt Stealing Attacks on In-the-Wild Prompts. In *IEEE Symposium on Security and Privacy*, pages 392–410, 2025.

[161] Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. In *International Conference on Learning Representations*, 2024.

[162] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[163] Knostic Team. Jailbreaking Social Engineering via Adversarial Digital Twins, 2024.

[164] Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models. In *Advances in Neural Information Processing Systems*, 2022.

[165] Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 103619–103651, 2024.

[166] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[167] Nikolaos Tsinganos, Panagiotis Fouliras, and Ioannis Mavridis. Leveraging Dialogue State Tracking for Zero-Shot Chat-Based Social Engineering Attack Recognition. *Applied Sciences*, 13, 2023.

[168] Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024.

[169] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[170] Yidan Wang, Yanan Cao, Yubing Ren, Fang Fang, Zheng Lin, and Binxing Fang. PIG: Privacy Jailbreak Attack on LLMs via Gradient-based Iterative In-Context Optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 9645–9660, July 2025.

[171] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[172] Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. Evaluating Copyright Takedown Methods for Language Models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[173] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[174] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, 2022.

[175] Jiankun Wei, Abdulrahman Abdulrazzag, Tianchen Zhang, Adel Muursepp, and Gururaj Saileshwar. Privacy risks of speculative decoding in large language models. *arXiv preprint arXiv:2411.01076*, 2024.

[176] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[177] Roy Weiss, Daniel Ayzenshteyn, and Yisroel Mirsky. What was your prompt? a remote keylogging attack on {AI} assistants. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 3367–3384, 2024.

[178] Guanlong Wu, Zheng Zhang, Yao Zhang, Weili Wang, Jianyu Niu, Ye Wu, and Yinqian Zhang. I know what you asked: Prompt leakage via kv-cache sharing in multi-tenant llm serving. In *Proceedings of the 2025 Network and Distributed System Security (NDSS) Symposium*, 2025.

[179] Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, et al. Differentially private synthetic data via foundation model apis 2: Text. In *International Conference on Machine Learning*, pages 54531–54560. PMLR, 2024.

[180] Qinge Xie, Karthik Ramakrishnan, and Frank Li. Evaluating privacy policies under modern privacy laws at scale: An {LLM-Based} automated approach. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 5797–5816, 2025.

[181] Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Gong, and Bhuwan Dhingra. ReCaLL: Membership Inference via Relative Conditional Log-Likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8671–8689, 2024.

[182] Wenpeng Xing, Zhonghao Qi, Yupeng Qin, Yilin Li, Caini Chang, Jiahui Yu, Changting Lin, Zhenzhen Xie, and Meng Han. MCP-Guard: A Defense Framework for Model Context Protocol Integrity in Large Language Model Applications. *arXiv preprint arXiv:2508.10991*, 2025.

[183] Jiacen Xu, Jack W. Stokes, Geoff McDonald, Xuesong Bai, David Marshall, Siyue Wang, Adith Swaminathan, and Zhou Li. AutoAttacker: A Large Language Model Guided System to Implement Automatic Cyber-attacks. 2024.

[184] Minrui Xu, Jiani Fan, Xinyu Huang, Conghao Zhou, Jiawen Kang, Dusit Niyato, Shiwen Mao, Zhu Han, Kwok-Yan Lam, et al. Forewarned is forearmed: A survey on large language model-based agents in autonomous cyberattacks. *arXiv preprint arXiv:2505.12786*, 2025.

[185] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study. In *Forty-first International Conference on Machine Learning*, 2024.

[186] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*, 2024.

[187] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering, 2024.

[188] Yifan Yang, Siqin Wang, Daoyang Li, Shuju Sun, and Qingyang Wu. GeoLocator: A Location-Integrated Large Multimodal Model for Inferring Geo-Privacy. *Applied Sciences*, 14(16), 2024.

[189] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023.

[190] Ren Yi, Octavian Suciu, Adria Gascon, Sarah Meiklejohn, Eugene Bagdasarian, and Marco Gruteser. Privacy Reasoning in Ambiguous Contexts. *arXiv preprint arXiv:2506.12241*, 2025.

[191] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially Private Fine-tuning of Language Models. In *International Conference on Learning Representations*, 2022.

[192] Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. A synthetic dataset for personal attribute inference. *Advances in Neural Information Processing Systems*, 37:120735–120779, 2024.

[193] Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu, Rachel Dias, Eric Wallace, Kai Xiao, Johannes Heidecke, et al. Trading inference-time compute for adversarial robustness. *arXiv preprint arXiv:2501.18841*, 2025.

[194] Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-Cost High-Power Membership Inference Attacks. In *International Conference on Machine Learning*, pages 58244–58282, 2024.

[195] Hang Zeng, Xiangyu Liu, Yong Hu, Chaoyue Niu, Fan Wu, Shaojie Tang, and Guihai Chen. Automated Privacy Information Annotation in Large Language Model Interactions. *arXiv preprint arXiv:2505.20910*, 2025.

[196] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual Memorization in Neural Language Models. In *Advances in Neural Information Processing Systems*, 2023.

[197] Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. Position: Membership Inference Attacks Cannot Prove That a Model was Trained on Your Data. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 333–345, 2025.

[198] Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-K%++: Improved Baseline for Pre-Training Data Detection from Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[199] Kaiyuan Zhang, Siyuan Cheng, Hanxi Guo, Yuetian Chen, Zian Su, Shengwei An, Yuntao Du, Charles Fleming, Ashish Kundu, Xiangyu Zhang, and Ninghui Li. SOFT: Selective Data Obfuscation for Protecting LLM Fine-tuning against Membership Inference Attacks. In *34st USENIX Security Symposium (USENIX Security 25)*, 2025.

[200] Kaiyuan Zhang, Zian Su, Pin-Yu Chen, Elisa Bertino, Xiangyu Zhang, and Ninghui Li. LLM Agents Should Employ Security Principles. *arXiv preprint arXiv:2505.24019*, 2025.

[201] Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Pretraining Data Detection for Large Language Models: A Divergence-based Calibration Method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5274, 2024.

[202] Yanzhe Zhang, Tao Yu, and Diyi Yang. Attacking vision-language computer agents via pop-ups. *arXiv preprint arXiv:2411.02391*, 2024.

[203] Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*, 2024.

[204] Zhiping Zhang, Bingcan Guo, and Tianshi Li. Privacy Leakage Overshadowed by Views of AI: A Study on Human Oversight of Privacy in Language Model Agent. *arXiv preprint arXiv:2411.01344*, 2024.

[205] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[206] Arman Zharmagambetov, Chuan Guo, Ivan Evtimov, Maya Pavlova, Ruslan Salakhutdinov, and Kamalika Chaudhuri. Agentdam: Privacy leakage evaluation for autonomous web agents. *arXiv preprint arXiv:2503.09780*, 2025.

[207] Xi Zheng, Zhuoyang Li, Xinning Gui, and Yuhan Luo. Customizing Emotional Support: How Do Individuals Construct and Interact With LLM-Powered Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.

[208] Xinyao Zheng, Husheng Han, Shangyi Shi, Qiyan Fang, Zidong Du, Xing Hu, and Qi Guo. Inputsnatch: Stealing input in llm services via timing side-channel attacks. *arXiv preprint arXiv:2411.18191*, 2024.

[209] Noé Zufferey, Sarah Abdelwahab Gaballah, Karola Marky, and Verena Zimmermann. "AI is from the devil." Behaviors and Concerns Toward Personal Data Sharing with LLM-based Conversational Agents. *Proceedings on Privacy Enhancing Technologies*, 2025(3):5–28, 2025.