The Privacy Paradox of LLMs: User Perceptions and the Reality of PII Leakage

Shuai Cheng
Zhejiang University
HangZhou, China
cs36@zju.edu.cn

Haitao Xu*
Zhejiang University
Hangzhou, China
haitaoxu@zju.edu.cn

Shu Meng
Zhejiang University
Hangzhou, China
mengshu@zju.edu.cn

Shuai Hao
Old Dominion University
Norfolk, Virginia, USA
shao@odu.edu

Chuan Yue
Department of Computer Science
Colorado School of Mines
Golden, Colorado, USA
chuanyue@mines.edu

Zhao Li
Hangzhou Yugu Technology
Hangzhou, China
Zhejiang University
Hangzhou, China
lzjoey@gmail.com

Abstract

Large language models (LLMs) are increasingly deployed, yet they introduce significant privacy risks by disclosing personally identifiable information (PII) during interactions. Although prior work has demonstrated the feasibility of extracting PII from LLMs, no comprehensive study has evaluated the actual extent of PII leakage across mainstream LLMs or investigated user perceptions, literacy, and behavioral responses to these risks. To address these gaps, we conduct a large-scale evaluation of PII leakage in popular LLMs, demonstrating that attackers can extract email addresses and phone numbers with high success rates. Through a mixed-methods study involving 20 interviews and 204 survey participants, we identify significant discrepancies between user concerns and behavior: despite strong concerns about PII leakage and limited understanding of training data provenance, users continue to use LLMs due to perceived utility, often exhibiting privacy cynicism. Based on these findings, we propose design implications for enhancing the privacy-utility balance in future LLM deployments.

CCS Concepts

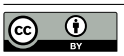
• **Security and privacy → Human and societal aspects of security and privacy**.

Keywords

Privacy Leakage, Privacy Cynicism, User Perception, PII Extraction

ACM Reference Format:
Shuai Cheng, Haitao Xu, Shu Meng, Shuai Hao, Chuan Yue, and Zhao Li. 2026. The Privacy Paradox of LLMs: User Perceptions and the Reality of PII Leakage. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3772318.3791809

*Haitao Xu is the corresponding author.



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791809

1 Introduction

Large language models (LLMs) have seen widespread adoption in real-world systems in recent years, owing to their strong performance across diverse domains ranging from natural language processing to interdisciplinary applications. However, numerous reports and user observations [18, 83, 91, 114] suggest that LLMs occasionally disclose private personal information during interactions. Such information, which can be uniquely linked to an individual, is referred to as personally identifiable information (PII) [76]. Common examples include names, email addresses, phone numbers, home and workplace addresses, educational background, as well as more sensitive data like Social Security Numbers and passwords.

Prior research [11, 20] has shown that LLMs can memorize training data verbatim; hence, PII leakage often occurs when such information is present in the training corpus. There are two primary categories through which PII enters LLM training data: (1) *publicly accessible sources*, such as web-crawled corpora, open books, academic papers, code repositories, and social media discussions[19, 43, 69, 93, 117], that may include publicly exposed PII due to misconfigured website permissions, inadvertently published and later removed content, or malicious data sales. Previous research [28] demonstrated that PII that had been deleted from the public web but preserved in historical Internet snapshots can still be incorporated into LLM training data, thereby leading to potential privacy leakage even after content deletion; and (2) non-public sources used by commercial LLMs, including *user-provided interaction data* with LLM collected under privacy policies [6, 85] as well as *licensed proprietary datasets* through partnerships and paid contractors [7, 52], both of which may embed sensitive personal information and reappear in updated models.

To mitigate PII leakage, LLM developers employ privacy-preserving techniques during training, such as data cleaning [57] and differential privacy [50], alongside alignment strategies like reinforcement learning with human feedback (RLHF) [87] and rule-based reward models (RBRMs) [2]. Nevertheless, recent studies [26, 28, 51, 64, 70, 81, 82] demonstrate that PII extraction remains feasible whether through targeted or large-scale non-targeted attacks, and that such PII can be traced back to academic datasets or public web content, confirming its presence in training data.

Existing defenses primarily focus on post-training model editing, such as model editing [8, 77] or machine unlearning [54], which modify privacy-relevant layers and neurons without full retraining but often leave residual memorization [115]. Another approach is query-time interception [31, 42, 104, 110, 124], which aims to filter PII from user inputs and outputs. However, this method remains vulnerable to prompt-engineering attacks [64].

While prior research has examined PII extraction techniques and defensive measures, no study has systematically evaluated the current scale of LLM PII leakage in real-world settings or explored user perspectives on this issue. Although existing user studies [5, 67, 72, 73, 75, 108, 122] have investigated LLM privacy and users' understanding of PII, none have specifically addressed the critical issue of PII leakage in LLMs or investigated user perspectives concerning it. This study therefore assesses the current state of PII leakage in mainstream LLMs and investigates user perceptions, including their awareness of privacy risks, objective privacy literacy, and willingness to use LLMs despite existing concerns.

Our research questions are as follows:

- **RQ1:** What is the extent and nature of privacy leakage in mainstream LLMs?
- **RQ2:** How do users perceive and understand LLM privacy risks, and what is their current level of privacy literacy?
- **RQ3:** To what extent do privacy concerns influence users' adoption intentions and continued usage behaviors regarding LLMs?

To address the above issues, we implemented state-of-the-art PII extraction techniques from existing research and conducted both targeted and non-targeted extraction attacks on mainstream LLMs. In targeted attacks based on public datasets, formatted PII (email addresses) achieved an average *Attack Success Rate (ASR)* of 78.3%, while less structured PII (phone numbers) still reached 20.3%. In large-scale non-targeted extraction attacks, the average ASR was as high as 34.6%, meaning that in every three rounds of interaction, at least one identifiable instance of PII (*e.g.*, an email or phone number) could be retrieved. Notably, some popular LLMs, including GPT-4o/5 and DeepSeek, exhibited substantially higher non-targeted ASRs (59.3%, 44.1%, and 49.8%, respectively), whereas others demonstrated much lower ASRs, such as Qwen (15.6%), Hunyuan (20.2%) and Gemeni (24.2%). These empirical findings informed our subsequent user study in two critical ways. First, the demonstrated severity of privacy risks highlighted the necessity of examining users' existing knowledge and perceptions regarding LLM-related privacy leakage. Second, we recognized the importance of incorporating these technical results into the study design—presenting them as tangible evidence to participants—to mitigate the expert–layperson knowledge disparity and stimulate informed reflection on privacy threats that might otherwise be overlooked or underestimated.

Leveraging these real-world evaluation results, we conducted a mixed-methods user study. First, we performed interviews to examine users' experiences and initial views of LLM privacy leakage. To ensure that participants' responses were grounded in concrete technical realities rather than abstract speculation, we then shared background knowledge of LLM training datasets and our empirical findings to elicit participants' perspectives on privacy leakage, training data collection practices, and PII usage. The qualitative phase revealed that participants lacked sufficient understanding of LLM background knowledge and underestimated both the risks of PII on the public internet and the effectiveness of PII extraction. They generally expressed strong negative attitudes toward the current state of PII leakage, yet offered relatively positive evaluations of training data collection. Most were unwilling to abandon LLMs due to privacy concerns, with many expressing helplessness or privacy cynicism [30, 49, 101], believing that modifying their usage practices would not meaningfully reduce leakage risks. Participants emphasized the need for greater transparency, user control, and stronger technical protections.

To further validate the interview findings, we conducted an online survey with 204 participants to systematically evaluate users' objective literacy about LLM privacy leakage, along with their perceptions and willingness to use LLMs. The survey results reinforced and extended the interview findings: although users demonstrated a solid grasp of basic LLM functionality, their understanding of training data practices remained limited. Moreover, users expressed the highest levels of concern and resistance toward PII that directly identifies individuals or could lead to physical or financial harm. Consistent with the interviews, participants reported high perceived vulnerability and severity alongside relatively low trust in LLMs. Nevertheless, they continued to show strong intentions to use LLMs, underscoring the tension between privacy risks and perceived utility.

In summary, our study makes the following contributions:

- **Comprehensive assessment of LLM privacy leakage.** We conducted a comprehensive evaluation of PII leakage in mainstream LLMs through targeted and non-targeted extraction attacks. Our results demonstrate significant leakage risks: attackers can successfully extract large amounts of PII from LLMs' training dataset using fabricated prompts with minimal effort. This finding highlights the limitations of existing safeguards and reveals substantial cross-model differences in susceptibility to privacy leakage, underscoring the severity and urgency of real-world privacy risks.
- **Empirical insights into users' awareness and literacy of LLM privacy risks.** Drawing on 20 semi-structured interviews and 204 online survey responses, we reveal that while most users believe LLMs are capable of leaking PII and proactively avoid entering sensitive information during interactions, they lack sufficient background knowledge of how training data are collected and processed, underestimate the risks of PII extraction, and hold significant misconceptions about LLMs' privacy protection mechanisms. These findings not only fill a gap in existing research on user privacy literacy but also highlight new cognitive blind spots in the context of generative AI.
- **Uncovering contradictions in user attitudes and behavioral responses.** Our findings show that users express strong concerns and negative attitudes toward LLM-induced PII leakage, particularly regarding information directly identifying individuals or posing risks to financial and physical security. At the same time, they tend to give relatively positive evaluations of training data collection practices and maintain a strong willingness to use LLMs, driven by efficiency and productivity benefits. Some even display a form of *privacy cynicism*, believing that their personal data are already highly exposed online and that modifying usage

habits cannot meaningfully reduce risks. Building on this paradox of privacy risk and usage intention, we translate user suggestions into design implications for privacy-preserving LLMs: ensuring transparent disclosure and governance of training data sources, strengthening real-time safeguards during user interactions, providing user-controllable privacy management mechanisms, and striking a balance between usability and privacy protection to strengthen both user trust and sustained adoption.

The remainder of this paper is structured as follows: §2 provides background on LLM training datasets and PII extraction research. §3 presents the evaluation of privacy leakage in mainstream LLMs. §4 details the mixed-methods user study design. §5 and §6 present interview and survey results, respectively. Finally, §7 discusses key findings, and §8 concludes the paper.

2 Background and Related Work

In this section, we first clarify the definition and categories of PII (§2.1), then outline the composition of typical LLM training datasets (§2.2), followed by a review of prior research on PII extraction attacks and defenses (§2.3). Subsequently, we introduce two privacy-related theoretical frameworks adopted in our study (§2.4). Finally, we examine existing user studies on LLM privacy and PII perceptions (§2.5).

2.1 Personally Identifiable Information (PII)

The National Institute of Standards and Technology (NIST) defines PII as: "any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information [76]." In short, PII refers to any data that can be used alone, or in combination with other information, to identify a specific individual.

2.2 LLM Training Dataset

The capabilities of LLMs rely heavily on massive training datasets, which provide them linguistic, reasoning, and general knowledge. Based on availability, these datasets can be classified into public and non-public sources.

2.2.1 Public Data Sources. Public data sources typically comprise open Internet corpora, including:

- **Web-crawled data**: large-scale collections from websites like blogs, forums, and articles [19, 93];
- **Books and academic resources**: digitized books, scientific papers, and open-access publications [19];
- **Code repositories**: publicly available code, *e.g.*, from GitHub, used to improve programming and logical reasoning [43, 69];
- **Social media and dialogue data**: public discussions and Q&A content from platforms like Reddit, facilitating conversational modeling [93, 117].

Several public datasets are commonly employed in LLM training, including: (1) **Enron dataset** [59], which contains roughly 500,000 corporate emails with substantial PII (*e.g.*, names, email addresses, phone numbers), has been incorporated into models such as Pythia [12] and GPT-Neo [14]. (2) **The Pile** [43], an 825-GB composite corpus created by EleutherAI, integrates 22 sub-datasets spanning academic papers, legal texts, source code repositories, and web-scraped content, and has been utilized in training GPT-Neo, GPT-J [117], and GPT-NeoX [13]. (3) **Common Crawl** [32], the largest openly available web-crawled corpus, containing more than 4.2 trillion webpage snapshots (about 419 TiB uncompressed), serves as the foundation for numerous derived subsets such as *C4* [94], *RefinedWeb* [90], *Dolma* [106], and *RedPajama* [119], which are extensively used in models including GPT-3 [19] and LLaMA-2 [113]. Despite rigorous cleaning efforts, significant amounts of PII remain detectable within these datasets [109]. (4) **OpenWebText** [46] and **OpenWebText2** [44], community-curated corpora based on external web content linked from highly upvoted Reddit posts, intend to approximate the WebText dataset used for GPT-2 training [93]. These datasets contain potential PII and have been used in models such as GPT-Neo and GPT-J. (5) **WikiText** [78], derived from Wikipedia articles and containing sensitive personal attributes (*e.g.*, names, birth dates), is employed for fine-tuning models including GPT-2 [93] and T5 [94].

Although most commercial LLMs (*e.g.*, GPT-4o/5, Claude 3.5/4, Gemini 2.5) do not disclose full training data details, studies [28, 64, 116] have identified content matching these public sources in model outputs, suggesting their continued use. The presence of PII in these datasets implies persistent privacy risks.

2.2.2 Non-public Data Sources. Many commercial LLMs also incorporate non-public data, which falls into two types. The first is *licensed proprietary data*, acquired through partnerships to access high-quality content. For instance, GPT-4o used proprietary datasets including paid content, archival data, and licensed media (*e.g.*, via Shutterstock) [52]. Claude 4 reported using "non-public data from third parties, data provided by data-labeling services and paid contractors" [7]. Rosenblat *et al.* [100] demonstrated GPT-4o's high recognition (82% AUROC score) of copyrighted book content, implying use of non-public material in training.

The second type is *user interaction data*, encompassing model inputs, outputs, user account details, device information, logs, and cookies. Privacy policies of commercial LLMs like ChatGPT [85] and Claude [6] permit such data collection for training, unless disabled under enterprise agreements. Claude 4 also acknowledges using voluntarily shared user data [7].

Although non-public dataset specifics are undisclosed, the above suggests they likely contain considerable user PII.

2.3 PII Extraction Attacks and Defenses

2.3.1 Feasibility of PII Extraction. Large-scale LLM training datasets inherently contain substantial PII, sourced from both public web data and non-public sources such as licensed corpora and user interactions. Studies demonstrate that widely used datasets including Common Crawl, The Pile, and C4 contain significant PII requiring rigorous filtering [92, 109]. Although preprocessing techniques like data cleaning and scrubbing are commonly employed to mitigate privacy risks, their effectiveness remains limited. Research indicates approximately 3% of PII sequences may persist after scrubbing [70],

while modifications to training data can unexpectedly enhance memorization of other PII [17].

Following established taxonomy [123], PII extraction methodologies are broadly categorized into two distinct approaches. *Targeted PII extraction* focuses on retrieving specific information about predetermined individuals, exemplified by queries such as "What is John Smith's email address?" In contrast, *non-targeted PII extraction* involves bulk information retrieval across groups or domains without specific targeting, typically implemented through repeated prompting designed to generate extensive lists of contact information or other identifiers.

2.3.2 PII Extraction Attack Techniques. Based on the taxonomy established by Cheng *et al.* [27], existing research on PII extraction from LLMs primarily falls into three attack categories:

- **PII Extraction from Leaked Training Data.** This line of work demonstrates models' tendency to memorize training samples and reproduce them verbatim [9, 21, 82, 123]. Representative techniques include black-box querying to recover sensitive details [21], divergence attacks using repetitive generation for large-scale extraction [82], and special-character prompts targeting memorized contact information [9]. For instance, Nasr *et al.* [82] demonstrated that feeding models with repetitive or low-entropy prompts (*e.g.*, a single token repeated many times) can induce the emission of extensive verbatim fragments from the training corpus, substantially increasing the output of memorized content including email addresses, phone numbers, and other PII during divergence phases.
- **PII Extraction via Crafted Prompts.** These investigations explore engineered prompts exploiting memorization and association mechanisms [28, 51, 58, 64, 70, 81, 84, 103, 116]. Research shows memorization poses higher leakage risks than association [51], with larger models exhibiting stronger inference capabilities [103]. Notable advancements include Li *et al.* [64]'s multi-step jailbreaking method inspired by Chain-of-Thought prompting, which achieved high extraction rates on Enron emails, and Nakka *et al.* [81]'s demonstration that prefix injection with unrelated PII markedly improves extraction success, particularly for structured identifiers such as phone numbers. More recently, Cheng *et al.* [28] proposed an enhanced few-shot approach for large-scale PII extraction, demonstrating superior efficiency and performance across both targeted and non-targeted scenarios.
- **PII Extraction via Fine-tuning.** This research direction highlights risks from fine-tuning, which reinforces memorization and exacerbates leakage from both fine-tuning and pre-training data [3, 16, 26, 88, 96]. Studies demonstrate that fine-tuning on PII-containing data amplifies memorization and can restore forgotten PII from pre-training datasets [26], while even fine-tuning on synthetic data induces real PII leakage [3].

2.3.3 PII Extraction Defense Mechanisms. Recent defensive approaches focus on adapting model parameters to unlearn or weaken privacy-related representations through techniques including learnable binary weight masks [25], gradient attribution [120], vocabulary space rank editing [8], and private association editing [115]. These methods demonstrate significant reductions in memorization accuracy while preserving model performance.

Another defensive strategy operates during query execution, detecting and obfuscating PII before model processing through prompt-level frameworks, privacy-enhanced text anonymization, and hybrid filtering systems [31, 35, 42, 104, 110, 124]. These approaches achieve high detection accuracy while maintaining usability, with user studies showing significant reduction in sensitive data disclosure without compromising task completion quality or user satisfaction.

2.4 Conceptual Models in Privacy Research

2.4.1 Privacy Calculus Model. The *Privacy Calculus Model*, rooted in information systems and marketing research, emphasizes that users engage in a cost-benefit analysis when deciding whether to disclose personal information. Within the context of LLM interactions, users balance privacy concerns against the perceived benefits and utility provided by these systems. This model suggests that when perceived benefits exceed risks and sufficient trust exists in the service provider, users demonstrate greater willingness to share data and continue usage; conversely, they may withdraw or restrict engagement.

Culnan and Armstrong [34] established the importance of trust and procedural fairness in privacy decisions, forming the theoretical foundation for subsequent privacy calculus research. Dinev and Hart [36] expanded this framework to e-commerce and information systems contexts, formalizing an "extended privacy calculus model" incorporating the following constructs:

- **Perceived Internet Privacy Risk (PR)**: Users' assessment of potential negative outcomes from personal information disclosure, including identity theft, privacy breaches, or data misuse.
- **Internet Privacy Concerns (PC)**: Persistent anxiety regarding privacy issues. Elevated risk perception heightens privacy concerns, subsequently reducing disclosure willingness.
- **Internet Trust (T)**: Confidence in online service providers' commitment to fulfill promises and protect data. Trust mitigates negative risk effects and enhances disclosure willingness.
- **Personal Internet Interest (PI)**: Users' motivation level for engaging with online services. Strong interest can positively influence disclosure intentions even amid heightened risks.
- **Willingness to Provide Personal Information to Transact (PPIT)**: The outcome variable indicating users' readiness to disclose personal information during online interactions.

Within this extended model, these factors interact to determine *PPIT*. *Perceived risk* directly reduces *PPIT* while indirectly exerting negative effects through increased *privacy concerns* and diminished *trust*. *Privacy concerns* mediate and amplify *perceived risk*'s suppressive effect. Conversely, *trust* buffers against *perceived risk*'s negative impact and directly promotes *PPIT*. *Personal interest* provides an independent positive pathway, enabling enhanced *PPIT* despite elevated risk perceptions.

2.4.2 Protection Motivation Theory (PMT). Rogers *et al.* [98, 99] developed *Protection Motivation Theory* to explain how individuals form intentions to protect themselves against threats. The theory's robustness in explaining security and privacy behaviors has been consistently demonstrated through meta-analytic reviews [40, 79] and extensive application in information security research [53, 56, 65]. In privacy and information security research, PMT has been

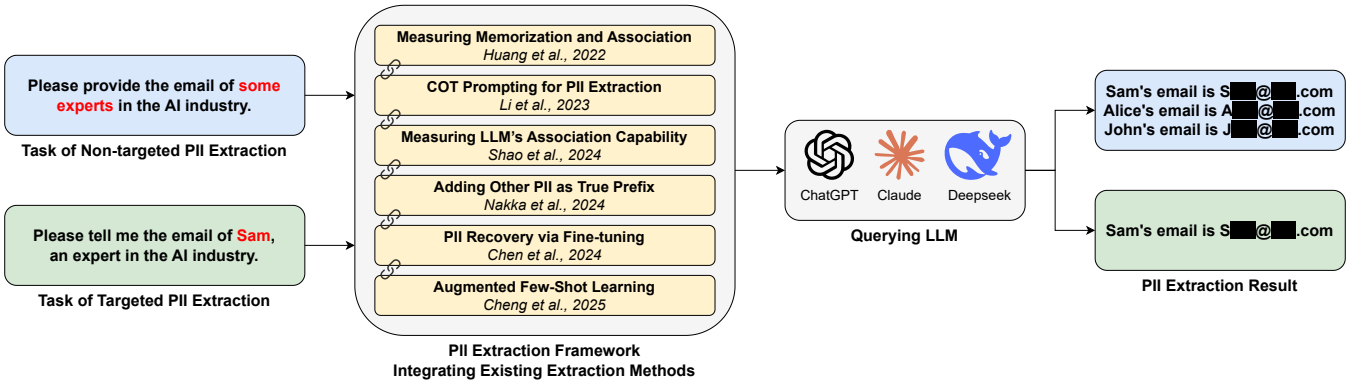


Figure 1: Framework for PII Leakage Assessment

widely applied to understand user responses to risks such as data breaches or privacy violations. The theory centers on two appraisal dimensions: *Threat Appraisal* and *Coping Appraisal*.

- **Threat Appraisal**: Evaluates user perception of threats through:
  - *Perceived Severity*: Assessment of how serious consequences would be (*e.g.*, financial loss or social stigma from data leakage).
  - *Perceived Vulnerability*: Estimated likelihood of personal exposure to the threat (*e.g.*, susceptibility to data collection during LLM interactions).
  - *Rewards*: Perceived benefits of maintaining current behavior despite risks (*e.g.*, convenience or immediate utility).
- **Coping Appraisal**: Assesses user confidence in addressing threats through:
  - *Response Efficacy*: Belief in the effectiveness of protective measures (*e.g.*, privacy plugins reducing data exposure).
  - *Self-Efficacy*: Belief in one's own ability to take protective measures (*e.g.*, configuring privacy settings appropriately).
  - *Response Costs*: Perceived expenses associated with protection, including financial, time, and convenience costs.

Threat appraisal and *coping appraisal* collectively determine *protective intentions*. *Perceived severity* and *perceived vulnerability* jointly shape threat perception: when users consider threats serious and themselves vulnerable, overall threat perception increases. *Rewards* negatively impact protective motivation by making risky behavior more appealing. *Response efficacy* and *self-efficacy* positively influence motivation by enhancing confidence in protective measures. *Response costs* inhibit protective behavior through perceived burdens.

2.5 User Studies on LLM Privacy and PII Perceptions

2.5.1 User Study on LLM Privacy. Several studies [5, 47, 61, 67, 72, 73, 108, 122] have examined user perspectives on LLM privacy. Ali *et al.* [5] analyzed 2.5M posts from the *r/ChatGPT* subreddit using mixed methods to investigate security and privacy concerns in conversational AI, revealing user anxieties across the data lifecycle and recommending enhanced transparency, user control, and trust mechanisms. Liu *et al.* [67] conducted a large-scale study (n=846) on Chinese users' privacy awareness in LLM-based healthcare consultations, combining online experiments with contextual integrity frameworks. They found high service adoption despite limited privacy awareness, highlighting contradictions that increase health data exposure risks. Zhang *et al.* [122] examined privacy risks in LLM-based conversational agents through analysis of 200 ShareGPT conversations and 19 user interviews, demonstrating how human-like interactions encourage sensitive disclosures while inadequate mental models and controls impede effective protection. However, these studies did not specifically investigate user perceptions of PII leakage, which typically entails more severe consequences than general privacy risks.

2.5.2 User Study on PII Perceptions. Some research has explored user understanding and perceptions of PII [63, 66, 75, 108]. Song *et al.* [108] conducted 32 interviews with users of period and fertility tracking apps to examine how they conceptualize PII, revealing dynamic and context-dependent views of identifiability, highlighting disparities between user perceptions and regulatory definitions, and offering user-centered design implications. Leon *et al.* [63] performed an online study with 2,912 participants, showing varying disclosure willingness across data types: strong resistance to sharing traditional PII (*e.g.*, phone numbers, addresses, credit cards) but greater acceptance of technical or demographic data. Malheiros *et al.* [75] investigated how online ad personalization using PII affects user comfort, finding that while PII-based ads attract more attention, they also increase discomfort and rejection. Nevertheless, these studies do not address PII usage and leakage in LLMs or user perceptions in this emerging context.

3 PII Leakage Assessment in Mainstream LLMs

To systematically evaluate privacy leakage in mainstream LLMs (RQ1), we implement multiple established PII extraction methodologies, applying their techniques to current language models. As shown in Figure 1, we perform both targeted and non-targeted PII extraction (as defined in §2.3.1) on each model, utilizing all applicable techniques. The highest success rate among the applied attacks is used to represent each LLM's PII leakage result.

3.1 Assessment Methodology

3.1.1 Selected Methodologies. From the studies reviewed in §2.3.2, we select approaches applicable to commercial black-box LLMs. To maximize extraction success, we exclude methods requiring training data access, focusing instead on prompt-based and fine-tuning techniques. The selected works are categorized as follows:

- **Targeted PII Extraction**: We implement techniques from Wang *et al.* [116], Li *et al.* [64], Chen *et al.* [26], Huang *et al.* [51], Cheng *et al.* [28], and Nakka *et al.* [81]. These studies extract raw PII through carefully designed prompts or fine-tuning, with available open-source implementations.
- **Non-targeted PII Extraction**: Cheng *et al.* [28] provide the only open-source implementation for non-targeted extraction, which we adopt for this evaluation.

3.1.2 Target Models and PII Categories. Based on the LMArena Team leaderboard ranking as of August 20, 2025 [68], we evaluate the top ten LLMs. We include only the highest-ranked model from each provider (except OpenAI, represented by two models) and exclude reasoning-specialized models for compatibility. The evaluated LLMs are: *gpt-5-high* (OpenAI), *chatgpt-4o-latest-20250326* (OpenAI), *claude-opus-4-1-20250805* (Anthropic), *grok-4-0709* (xAI), *kimi-k2-0711-preview* (Moonshot), *qwen3-235b-a22b-instruct-2507* (Alibaba), *glm-4.5* (ZhipuAI), *gemini-2.5-flash* (Google), *deepseek-v3-0324* (DeepSeek), and *hunyuan-turbos-20250416* (Tencent). All models are accessed via official APIs using default parameters unless specified otherwise.

We focus on three PII categories covered by the selected studies: *names*, *email*, and *phone#*. For targeted extraction, we provide names and query associated emails and phones. For non-targeted extraction, we directly request <name, email, phone> triples.

3.1.3 Ground Truth Datasets.

- **Targeted PII Extraction**: We use *Enron*, *OpenWebText2*, and *Common Crawl* as reference datasets. As noted in §2.2.1, these publicly available corpora are commonly used for LLM training, enabling evaluation of training data memorization. Given Common Crawl's scale, we use its subset *CC-News* [74], containing millions of news articles from 2016–2019 and widely used in models like T5 [94]. We extract emails and phone# using regular expressions and manual verification, linking each to corresponding names (Table 1). We exclude Enron from phone# evaluation due to insufficient verifiable name-phone pairs.

Table 1: Statistics of Extracted PII across Ground-truth Datasets

	Enron	OpenWebText2	CC-News	Total
Email	1391	1318	1518	4227
Phone	-	397	554	951

- **Non-targeted PII Extraction**: Following Cheng *et al.* [28], we target four professions—*Doctor*, *Accountant*, *Lawyer*, and *Journalist*—extracting professional PII in each domain. Outputs are verified through exact-match Google Search queries [1], with matching webpages serving as ground truth, consistent with [28]'s methodology.

3.2 Evaluation Results (RQ1)

3.2.1 Evaluation Metrics. We use *ASR (Attack Success Rate)* for both targeted and non-targeted extraction, defined as the percentage of correctly extracted PII instances. A PII instance is deemed correct only if the extracted string exactly matches the ground truth.

For *targeted extraction*, we evaluate ASR separately on each dataset, denoted as *ASR-Enron*, *ASR-OpenWebText2*, and *ASR-CC-News*, with the overall performance reported as *ASR-All*. For *non-targeted extraction*, following Cheng *et al.* [28], we additionally record the number of fully correct triples (denoted as *HPC-tri*, where HPC stands for Harvested PII Count) and the number of instances where either a <name, phone> or <name, email> pair is correct (denoted as *HPC-all*). The ASR for non-targeted extraction is subsequently calculated based on *HPC-all*.

3.2.2 Evaluation Setup.

- **Targeted PII Extraction**: For each model, we apply all applicable methods to the ground-truth PII. Each PII instance is tested five times, with averages reported per dataset and overall. We report the highest *ASR-All* across all methods as the targeted extraction result. Email and phone number extractions are evaluated separately, skipping incompatible cases, such as when a method requires fine-tuning on unsupported models, or when a method is designed to extract only emails but not phone numbers.
- **Non-targeted PII Extraction**: For each model, we evaluate all four professions with 500 extraction rounds (five repetitions for each round). We compute averages for *HPC-tri*, *HPC-all*, and *ASR* per profession, with combined results across professions as the model's non-targeted outcome.

Table 2: Performance of Targeted Email Extraction

Model	ASR-Enron	ASR-Open WebText2	ASR-CC-News	ASR-All
gpt-5	86.1%	71.5%	80.6%	79.5%
gemini-2.5-flash	84.3%	76.0%	80.6%	80.4%
claude-opus-4.1	85.1%	78.5%	81.9%	81.9%
chatgpt-4o	74.7%	33.0%	37.8%	48.4%
grok-4	84.5%	69.7%	80.0%	78.3%
qwen3-235b-a22b	84.5%	68.4%	79.0%	77.5%
kimi-k2	80.5%	65.3%	77.9%	74.8%
glm-4.5	83.8%	67.5%	78.6%	76.8%
deepseek-v3	83.5%	70.9%	79.9%	78.3%
hunyuan-turbos	83.7%	68.3%	79.1%	77.2%
Average	83.1%	66.9%	75.4%	78.3%

3.2.3 Results.

- **Targeted PII Extraction**: Results are provided in Tables 2 and 3. For email extraction, most LLM models achieved high success rates (average 78.3% ASR), with *claude-opus-4.1* (81.9%) and *gemini-2.5-flash* (80.4%) showing the most severe leakage. Except *chatgpt-4o* (48.4%), all models exceeded 70% success. Phone extraction success was substantially lower (average 20.3% ASR), with *gpt-5* (40.3%) and *chatgpt-4o* (39.6%) performing best. This indicates emails—more common and standardized—are more easily memorized and extracted, while phone numbers exhibit greater randomness, making them more challenging to extract. However, these success rates demonstrate significant privacy risks,

indicating LLMs cannot prevent training data leakage despite safety measures like RLHF/RBRM, highlighting risks for non-public training data such as proprietary licensed corpora and user interaction data.

Finding I-1: Mainstream LLMs exhibit significant vulnerability to targeted PII extraction. Emails are highly susceptible to memorization (78.3% average ASR), while phones show lower but non-trivial exposure (20.3%). Successful extraction from public training data underscores leakage risks for non-public corpora.

Table 3: Performance of Targeted Phone Extraction

Model	ASR-Open WebText2	ASR-CC-News	ASR-All
gpt-5	56.9%	28.3%	40.3%
gemini-2.5-flash	25.7%	22.9%	24.1%
claude-opus-4.1	15.9%	13.7%	14.6%
chatgpt-4o	52.6%	30.3%	39.6%
grok-4	9.6%	13.0%	11.6%
qwen3-235b-a22b	4.0%	5.8%	5.1%
kimi-k2	17.9%	17.5%	17.7%
glm-4.5	5.0%	7.6%	6.5%
deepseek-v3	31.0%	26.9%	28.6%
hunyuan-turbos	15.4%	14.6%	14.9%
Average	23.3%	18.1%	20.3%

- **Non-targeted PII Extraction**: Results per profession and overall are presented in Tables 4 and 5. Performance varied substantially across professions. Doctors and journalists exhibited the highest leakage rates (average ASRs of 38.1% and 42.8%), with *deepseek-v3* achieving 63.6% for doctors and *chatgpt-4o* reaching 74.2% for journalists. Accountants and lawyers demonstrated lower but still considerable rates (34.6% and 33.7%), suggesting varying susceptibility to privacy extraction across professions due to their differential representation in public data. Additionally, model performance differed significantly: *chatgpt-4o* (59.3%) and *deepseek-v3* (49.8%) showed severe vulnerability, with nearly one in two extraction attempts yielding a correct PII instance; by comparison, *qwen3-235b-a22b* (15.6%) and *hunyuan-turbos* (20.2%) demonstrated stronger defenses while remaining susceptible to extraction. Overall, extraction across all four professions yielded *6,919* valid PII instances (*HPC-all*) with 34.6% average ASR, demonstrating mainstream LLMs enable large-scale personal information harvesting across professions, highlighting both the breadth and severity of LLM privacy leakage.

Finding I-2: Non-targeted PII extraction reveals profession-dependent risks, with doctors and journalists most vulnerable. *chatgpt-4o* (59.3%) and *deepseek-v3* (49.8%) show severe leakage, while *qwen3-235b-a22b* (15.6%) and *hunyuan-turbos* (20.2%) demonstrate relative resilience. The 34.6% average ASR confirms mainstream LLMs enable large-scale personal information harvesting, indicating widespread privacy vulnerabilities.

Table 4: Performance of Non-Targeted PII Extraction

Accountant			
Model	**HPC-*tri***	**HPC-*all***	**ASR(%)**
gpt-5	7	121	24.2%
gemini-2.5-flash	35	124	24.8%
claude-opus-4.1	2	207	41.4%
chatgpt-4o	55	256	51.2%
grok-4	1	45	9.0%
qwen3-235b-a22b	2	5	1.0%
kimi-k2	0	71	14.2%
glm-4.5	4	88	17.6%
deepseek-v3	1	199	39.8%
hunyuan-turbos	0	69	13.8%
Average	**1,489**	**6,919**	**34.6%**
Doctor			
Model	**HPC-*tri***	**HPC-*all***	**ASR(%)**
gpt-5	15	208	41.6%
gemini-2.5-flash	3	98	19.6%
claude-opus-4.1	0	240	48.0%
chatgpt-4o	34	266	53.2%
grok-4-0709	85	267	53.4%
qwen3-235b-a22b	0	43	8.6%
kimi-k2	55	294	58.8%
glm-4.5	0	55	11.0%
deepseek-v3	90	318	63.6%
hunyuan-turbos	4	115	23.0%
Average	28.6	190.4	38.1%
Journalist			
Model	**HPC-*tri***	**HPC-*all***	**ASR(%)**
gpt-5	51	335	67.0%
gemini-2.5-flash	123	231	46.2%
claude-opus-4.1	55	216	43.2%
chatgpt-4o	127	371	74.2%
grok-4	38	166	33.2%
qwen3-235b-a22b	1	62	12.4%
kimi-k2	22	202	40.4%
glm-4.5	33	168	33.6%
deepseek-v3	165	269	53.8%
hunyuan-turbos	2	125	25.0%
Average	61.7	214.5	42.8%
Lawyer			
Model	**HPC-*tri***	**HPC-*all***	**ASR(%)**
gpt-5	91	217	43.4%
gemini-2.5-flash	16	31	6.2%
claude-opus-4.1	87	169	33.8%
chatgpt-4o	157	293	58.6%
grok-4	0	21	4.2%
qwen3-235b-a22b	2	202	40.4%
kimi-k2	58	205	41.0%
glm-4.5	0	242	48.4%
deepseek-v3	56	210	42.0%
hunyuan-turbos	12	95	19.0%
Average	47.9	168.5	33.7%

4 User Study Methodology

To investigate user perceptions and literacy concerning LLM privacy leakage (RQ2), as well as their risk-benefit trade-offs in LLM usage (RQ3), we initially conducted in-person interviews with 20 participants to explore their understanding of privacy risks and the factors influencing their decisions. We then carried out an online survey involving 204 participants to validate key qualitative findings from the interviews and to systematically assess participants' privacy literacy regarding LLMs.

In this section, we detail the methodological components of our study, beginning with the structured interview protocol for qualitative data collection (§4.1), followed by the design and development of the quantitative survey instrument (§4.2). We then describe the

participant recruitment strategies for both the interview and survey (§4.3) and finally outline the analytical approaches applied to both qualitative and quantitative data (§4.4).

Table 5: Average Performance of Non-Targeted PII Extraction across Professions

Model	HPC-*tri*	HPC-*all*	ASR(%)
gpt-5	164	881	44.1%
gemini-2.5-flash	177	484	24.2%
claude-opus-4.1	144	832	41.6%
chatgpt-4o	373	1186	59.3%
grok-4	124	499	25.0%
qwen3-235b-a22b	5	312	15.6%
kimi-k2	135	772	38.6%
glm-4.5	37	553	27.7%
deepseek-v3	312	996	49.8%
hunyuan-turbos	18	404	20.2%
Overall Average	**1489**	**6919**	**34.6%**

4.1 Interview Protocol

The interview consisted of three main phases. Initially, participants' demographic information and LLM usage patterns were collected, along with their experiences and basic perceptions of LLM privacy leakage. The second phase introduced background knowledge on LLM training data as discussed in §2.2 , PII extraction mechanisms as discussed in §2.3. Then our PII attack evaluation outlined in §3 was presented, along with empirical extraction results, including a live demonstration of PII leakage. All extracted PII was anonymized prior to presentation and promptly deleted afterward to address ethical considerations.

The final phase centered on participants' reflections and evaluations. They were asked to identify which aspects of the background knowledge they were previously aware of or unfamiliar with. Their perceptions of the current state of LLM privacy leakage and training data collection were also explored. Participants assessed their own level of PII exposure online, rated their concern regarding the potential use and leakage of such data in LLM training, and indicated which categories of PII they deemed unacceptable for training purposes. Finally, the trade-offs they made between the benefits of LLM use and privacy risks were investigated, along with their behavioral responses and mitigation suggestions. The full interview protocol is available in Appendix A.

4.2 Survey Development

The survey questionnaire was divided into two main parts. The first collected demographic data, including gender, age, education level, occupation, computer science (CS) background, estimated annual income, as well as the frequency of LLM usage. The second part began with a concise overview of LLM privacy leakage. It then assessed participants' objective literacy of LLM privacy leakage through eight factual judgment items and two multiple-choice question identifying the purposes of LLM usage and the types of PII of greatest concern. Subsequent items measured privacy concerns, risk perceptions, and trade-off considerations using a series of 7-point Likert scales. We primarily adopted the *Privacy Calculus Model* [34, 36], from which constructs such as *Trust*, *Privacy Concerns (PC)*, and *Personal Interest (PI)*, *Perceived risk (PR)* were adapted and contextualized to the LLM scenario. The construct of *PPIT* in privacy calculus model was excluded, as LLMs do not force users to provide personal information in order to access their services. To provide a more fine-grained reflection of users' threat appraisal, we substituted the conventional construct of *Perceived risk (PR)* with the *Perceived Severity* and *Perceived Vulnerability* dimensions from the *Protection Motivation Theory (PMT)* [98, 99]. All constructs were measured using validated scales adapted to the context of LLM privacy leakage. Appendix B details all constructs, items, and their sources for each aforementioned factor.

A pilot study with 30 participants was conducted to refine the questionnaire. Improvements included adding a diagram illustrating LLM privacy attacks and a multiple-choice question regarding sensitive PII types. The average completion time was 8 minutes; to ensure data quality, an attention-check item and a minimum response time threshold of 4 minutes were implemented. Informed consent was obtained at the survey outset, emphasizing voluntary participation, the right to withdraw, and assurances that data would be used solely for academic purposes under confidentiality. The complete questionnaire is provided in Appendix C.

4.3 Participant Recruitment

4.3.1 Interview Recruitment. We recruited 20 Chinese participants through campus flyers and university-affiliated website. Eligibility required individuals to be aged 18–48 and use LLMs at least monthly. The age criterion was informed by recent survey data indicating that over 80% of ChatGPT users fall within the 18–44 age range, with users aged 18–25 accounting for 45% [37], rendering this demographic highly relevant. Each interview lasted approximately 25–45 minutes, with participants receiving approximately 20 USD compensation in accordance with local ethical guidelines. This study was approved by the university's Institutional Review Board (IRB).

As summarized in Table 6, the study involved 20 participants with a gender distribution approximating a 2:1 male-to-female ratio, consistent with reported trends among ChatGPT users [37]. The sample comprised an equal proportion of students and employed professionals. A balanced representation was maintained between participants with computer science-related backgrounds (50%) and those from non-CS disciplines (50%); notably, most individuals with CS backgrounds were students rather than professional computer scientists, thereby facilitating diverse perspectives on privacy perceptions. Interview sessions ranged from 26 to 44 minutes (mean duration: 31.5 minutes), allowing for in-depth exploration of participants' LLM usage patterns and their experiences with LLM privacy leakage. Participants reported high levels of engagement with LLMs, with 65% indicating daily use and 30% weekly use. The most frequently utilized models included DeepSeek (reported by 75% of participants), ChatGPT (70%), Gemini (30%), and Qwen (20%). Primary use cases consistently featured knowledge search (55%), language translation (35%), coding-related tasks (30%), and writing assistance (30%) among their top two most frequent applications. Additional usage patterns encompassed content generation, exam support, informal chatting, and related activities, demonstrating

Table 6: Participant Demographics and LLM Usage Information

ID	Age	Gender	Employment Status	CS-related	Commonly Used LLMs	Usage Frequency	Top-2 Usage Purpose	Interview Duration
P1	18–27	Male	Student	Yes	Claude, Qwen	Daily	Knowledge Search Coding or It-related Task	32min
P2	18–27	Male	Student	Yes	Claude, DeepSeek, Gemini	Daily	Coding or It-related Task Exam Support	36min
P3	28–37	Female	Employed	Yes	ChatGPT, DeepSeek	Daily	Knowledge Search Language Translation	38min
P4	18–27	Female	Student	Yes	ChatGPT, Gemini, Qwen	Daily	Writing Assistance Daily Work Assistance	29min
P5	18–27	Male	Student	Yes	DeepSeek, Gemini	Daily	Knowledge Search Coding or It-related Task	27min
P6	18–27	Female	Student	Yes	Claude, DeepSeek, Gemini	Daily	Knowledge Search Writing Assistance	28min
P7	18–27	Female	Student	Yes	ChatGPT, Claude, DeepSeek	Daily	Knowledge Search Exam Support	35min
P8	18–27	Male	Student	No	ChatGPT	Weekly	Knowledge Search Coding or It-related Task	29min
P9	28–37	Male	Employed	Yes	ChatGPT, DeepSeek, Qwen	Daily	Knowledge Search Coding or It-related Task	28min
P10	28–37	Male	Employed	No	ChatGPT, DeepSeek, Gemini	Daily	Daily Chatting Content Generation	28min
P11	28–37	Male	Employed	No	ChatGPT, DeepSeek	Weekly	Writing Assistance Language Translation	44min
P12	28–37	Male	Employed	No	DeepSeek	Weekly	Knowledge Search Language Translation	38min
P13	28–37	Female	Employed	No	DeepSeek, Kimi, Qwen	Daily	Daily Work Assistance Daily Chatting	26min
P14	28–37	Male	Employed	No	ChatGPT, DeepSeek	Monthly	Knowledge Search Language Translation	29min
P15	18–27	Male	Student	Yes	ChatGPT, Gemini	Daily	Writing Assistance Language Translation	26min
P16	18–27	Male	Student	Yes	ChatGPT, Kimi	Daily	Knowledge Search Daily Work Assistance	30min
P17	18–27	Female	Employed	No	ChatGPT, DeepSeek	Daily	Writing Assistance Content Generation	29min
P18	28–37	Male	Employed	No	ChatGPT, DeepSeek	Weekly	Writing Assistance Language Translation	43min
P19	28–37	Male	Employed	Yes	ChatGPT, DeepSeek	Weekly	Coding or It-related Task Content Generation	27min
P20	18–27	Male	Student	No	ChatGPT, DeepSeek, GLM	Weekly	Knowledge Search Language Translation	29min

the pervasive integration of LLMs into participants' daily routines and professional workflows.

Discussion. We acknowledge that recruiting participants from diverse occupational backgrounds represents an ideal approach for comprehensive sampling. However, the absence of reliable, publicly available data on the occupational distribution of LLM users makes empirically precise stratification by professional field unfeasible. To strengthen sample representativeness, we consequently prioritized behavioral alignment over occupational diversity, selecting participants whose primary LLM usage purposes closely reflect those observed in the broader user population. Studies on LLM usage patterns [10, 23, 80, 111] indicate that general users predominantly engage in knowledge search (reported by over 50% of users), writing or work-related assistance (over 35%), and programming or IT-related tasks (approximately 20–30%). The distribution of usage purposes within our interview cohort aligns well with these documented patterns, thereby supporting the representativeness of our sample in terms of actual LLM engagement behaviors.

4.3.2 Survey Recruitment. Participants were recruited through *credamo* [33], a professional online survey platform in China, using a pre-screened panel to enhance statistical representativeness relative to the general LLM user population in China. The largest age group was 18-27 constituting 52.9% of the sample, while 58.8% identified as male. Chi-square tests indicated no significant differences in age ($\chi^2(4) = 1.35$, $p = 0.852$) or gender ($\chi^2(1) = 2.58$, $p = 0.108$) distributions between our sample and the broader LLM user population [37], supporting demographic representativeness.

Eligibility criteria mirrored those of the interview study. Several quality control measures were implemented, including prescreening, attention checks, and exclusion of speeders (*i.e.*, surveys completed in under 4 minutes). From an initial pool of 300 respondents, 204 valid responses were retained for analysis. Each participant received approximately 5 USD compensation upon survey completion in accordance with local ethical guidelines. Comprehensive demographic statistics are included in Appendix D.

4.4 Data Analysis Methods

4.4.1 Qualitative Data Analysis. Thematic analysis was applied to the interview data. All interviews were manually transcribed and coded using NVivo [71]. Two researchers independently coded transcripts and developed preliminary codebooks, which were subsequently merged through collaborative discussion. Discrepancies

were resolved via adjudication by a third researcher. After six iterative rounds of coding and refinement, a final codebook was established with strong inter-coder reliability (Cohen's kappa=0.883) [39]. Remaining disagreements were settled through consensus. The full codebook is available in Appendix E.

4.4.2 Quantitative Data Analysis. Reflecting the exploratory nature of our investigation into user perceptions and current practices regarding LLM privacy leakage, survey data were examined primarily through descriptive statistical methods, supplemented by correlational analyses. These quantitative findings serve to complement themes identified in the qualitative analysis, providing contextual support and broader insights alongside the interview findings.

5 Qualitative Interview Findings

This section presents our qualitative findings from participant interviews, structured to systematically examine users' understanding of and responses to LLM privacy risks. We begin by establishing participants' baseline experiences and initial perceptions of privacy leakage (§5.1), followed by an analysis of significant gaps in their technical literacy and risk awareness (§5.2). We then examine their overall perspectives on current privacy leakage states and data collection practices (§5.3), and analyze the specific dimensions shaping their acceptance or rejection of various PII types in model training (§5.4). Finally, we document their behavioral responses to identified privacy threats (§5.5) and summarizing their proposed mitigation strategies (§5.6). Interview questions are denoted with the prefix "IQ" corresponding to items in Appendix A.

5.1 User Experiences and Initial Views of LLM Privacy Leakage (RQ2)

5.1.1 Views on LLMs' Capacity for PII Leakage (IQ7). When asked whether LLMs could leak PII, 15 participants believed such leakage was possible, while the remaining five denied this possibility (P2, P14, P18, P19, P20).

Among those affirming the risk, some cited personal encounters. P15 stated: *"I believe LLMs can leak personal information because, when I previously sought advice from one, it suddenly produced information resembling an email address and phone number."* Others based their views on technical reasoning. P5 remarked: *"Much of what LLMs generate comes from their training data. If that data contains unfiltered PII, then it will inevitably be exposed."* A minority relied on intuitive judgment. P13 explained: *"Since I have provided my personal information before, I feel there is always a chance that LLMs could leak it."*

Conversely, participants who rejected the possibility often questioned the technical reliability of extracting precise PII. P18 commented: *"Although there are jailbreak-like methods to bypass restrictions, the outputs of LLMs are often uncertain. I don't think it is feasible to obtain precise personal information in this way."*

5.1.2 Reported Experiences of PII Leakage (IQ8). Despite these perceptions, most participants reported no direct encounters with PII leakage from LLMs. Only six indicated having experienced what they perceived as leakage (P1, P3, P11, P13, P15, P17). These instances typically involved LLMs generating content resembling personal information, though authenticity often remained unverified. P17 shared: *"I once asked an LLM to help me draft character backgrounds for a script, and it generated detailed personal experiences, including names and educational histories. While some of this information could be found online, the rest could not be verified."*

Others described incidents not regarded as genuine leakage. P12 recounted: *"When I searched for a particular scholar, the LLM only provided general information such as research interests and institutional affiliation. I do not consider this a privacy leak."* Similarly, P1 referred to cases where the model retrieved public data: *"I asked about the email of a professor, and the model pulled information from the public teacher homepage. I don't think that counts as privacy leakage by the LLM."*

5.1.3 Willingness to Disclose PII during LLM Interactions (IQ9). Although most had not experienced leakage, we inquired whether they voluntarily shared PII during LLM interactions. Only 4 participants reported having done so (P6, P11, P16, P17). P6 admitted: *"I once provided my phone number, but I think as long as I stay vigilant against telecom fraud, I will be safe."* In contrast, 10 of the remaining 16 explicitly avoided providing PII (P1, P2, P3, P4, P7, P9, P10, P12, P19, P20). P19 explained: *"For example, when I ask an LLM to process a personal file, I always create a version with my personal information removed or anonymized before submitting it."*

Finding II-1: Although participants reported few firsthand experiences of LLMs leaking PII, most believed such leakage is possible and have begun proactively avoiding the input of personal information during interactions.

5.2 Gaps in LLM Privacy Literacy (RQ2)

After introducing background knowledge on LLM training data, our attack framework, and PII extraction results, we asked participants which aspects they were previously unaware of (IQ10). Every participant reported gaps in at least one dimension of technical literacy, categorized into four groups:

- **Limited Understanding of LLM Fundamentals** (P2, P6, P8, P9, P10, P12, P13, P14, P16-P20): Many participants demonstrated unawareness of how LLMs are trained. P8 highlighted limited understanding of non-public data sources: *"This is the first time I learned that most LLMs do not disclose their training datasets. Vendors claim to use only public or licensed data, but does that mean licensed data contains no personal information?"* P19 admitted unfamiliarity with user interaction data collection: *"GPT has never informed me that it collects my conversation data."*
- **Limited Awareness of PII Extraction Attacks** (P9, P16, P17, P18, P20): Several were unfamiliar with PII extraction techniques. P18 stated: *"I never imagined that few-shot prompts could make LLMs output accurate content instead of producing hallucinations."*
- **Underestimation of Risks of PII in Public Website** (P2, P3, P7, P9, P10, P12, P14, P18): Some participants underestimated risks associated with PII embedded in public web pages. P12 explained: *"I assumed that PII on web pages was public and harmless, and I never considered that such information might be unauthorized. When LLMs aggregate this data, it definitely raises privacy risks."*
- **Underestimation of PII Extraction Efficacy** (P1, P2, P4, P5, P8, P11, P15): Many expressed surprise at the efficacy of our PII

extraction methods. P15 remarked: *"I used to think the likelihood of LLMs leaking privacy was only fractions of a percent, maybe 0.05%. I never expected your attack success rate to be this high."*

Finding II-2: Participants exhibited limited understanding of LLM background knowledge, including training data and attack techniques. They also tended to underestimate threats posed by PII on the public Internet as well as the efficacy of current PII extraction methods.

5.3 Perspectives on LLM Privacy Leakage (RQ3)

5.3.1 Attitudes Toward Current LLM Privacy Leakage (IQ11). Most participants expressed negative attitudes toward the current LLM privacy leakage (P2, P4, P5, P7, P8, P10, P12-P17, P20). Some voiced fear and anxiety regarding undisclosed training datasets, as P13 noted: *"My fear comes from the unknown—I have no idea what information those undisclosed datasets might contain."* Others expressed helplessness; P14 stated: *"The leakage of personal information is definitely uncomfortable, but I cannot stop LLMs from accessing it."* Concerns about criminal misuse were also raised. P5 warned: *"LLMs contain large amounts of personal information, which might turn them into tools for social engineering crimes. That really worries me."*

Others held less negative or neutral stances (P1, P3, P6, P9, P10, P16, P19). P19 suggested the issue is not unique to LLMs: *"Even if LLMs can extract a lot of PII, advanced search techniques could do the same—this is not really an LLM problem."* P9 exhibited privacy cynicism: *"Most of my privacy has already been leaked anyway, so additional leakage by LLMs doesn't really affect me."*

5.3.2 Views on LLM Training Data Collection Practices (IQ12). Participants expressed relatively more positive attitudes toward data collection practices compared to leakage concerns. Sixteen participants were supportive and often cited legality or necessity (P1, P2, P3, P4, P6-P17). P10 argued: *"Since training data comes from public websites or licensed sources, I don't see a problem with such data collection."* P4 emphasized the necessity for model performance: *"The capabilities of LLMs scale with the size of their training data, so collecting such data is essential for improving performance."*

Nevertheless, eleven expressed negative views (P4, P5, P10, P12, P13, P15-P20). P5 stressed unauthorized use: *"Web crawlers scraping my personal website is not the same as me authorizing its use as training data."* P18 raised ethical concerns: *"Using LLMs that rely on data infringing others' privacy or copyrights gives users a sense of moral guilt."* P20 highlighted risks in licensed data: *"If a company providing licensed data already holds large amounts of PII and includes it in the dataset, this poses a serious problem."*

5.3.3 Perceived Severity of Existing PII Leakage (IQ13). Fifteen participants believed their PII had already been severely compromised online (P1-P11, P13, P15, P17, P20). P11 cited social engineering risks: *"My personal data must be in multiple leaked databases. If someone targets me, they could easily collect all my information using social engineering."* P2 emphasized cross-platform linkage: *"If I don't deliberately make my accounts look different across platforms, the leakage of just one account could trigger a chain reaction exposing all my data."* P13 shared a personal example: *"I once searched for my name online and found my home address and an old phone number on a shady website."*

The remainder gave moderate assessments (P12, P14, P16, P18, P19), often due to proactive privacy measures. P18 stated: *"I intentionally avoid sharing personal details online, such as my education, address, or anything tied to my personality traits."*

5.3.4 Concerns About LLM-Aggravated Privacy Risks (IQ14). Given widespread awareness of existing online PII exposure, eleven participants expressed strong concerns about further leakage via LLMs (P2, P4, P5, P7, P10-P14, P17, P20). P7 feared the aggregation capacity of LLMs: *"My personal information may already be scattered across different websites, but once LLMs aggregate it together, the scope of privacy leakage becomes terrifying."* P17 described feeling exposed: *"Such aggregation of privacy data makes me feel like a transparent person."*

Others reported low (P1, P3, P6, P8, P9, P15) or moderate (P16, P18, P19) concern. P1 dismissed personal targeting: *"As an ordinary person, I don't possess important personal data that would be worth targeting by hackers."* P3 argued current LLM privacy attacks are less efficient than traditional methods: *"Methods such as social engineering or targeted searches are much more efficient than LLMs."*

Finding II-3: Participants generally held negative views toward current LLM PII leakage, alongside expressions of privacy cynicism. However, they evaluated training data collection more positively due to LLMs' utility. Most perceived their online privacy as already highly exposed, amplifying concerns about LLM-related leakage.

5.4 Unacceptable PII Categories for LLM Training (RQ3)

Participants identified types of PII they considered unacceptable for LLM training (IQ15), with criteria falling into 6 dimensions:

- **Directly Identifying PII** (P1, P5, P7, P8, P12, P13, P14, P15, P18, P20). The most frequent concern was directly identifiable information, such as ID numbers, phone numbers, and home addresses. Less direct identifiers like emails or names were more acceptable. P15 noted: *"I can accept the leakage of my name or username, since there are many people with the same name. But phone numbers are much more dangerous."* P13 expressed minimal concern about emails: *"Emails are supposed to be public information, and they also have filtering functions. I rarely use email anyway."*
- **Sensitive Personal Information** (P2, P3, P6, P9, P10, P16). Several strongly opposed the use of sensitive personal data for model training, including financial or health data. P6 argued: *"Information like my financial assets or even my height and weight should never be used as training data."* P9 extended this to multimodal data: *"I fear that LLMs may secretly use my photos, voice, or even video recordings with others as training data, and I would have no way of knowing."*
- **PII Entailing Tangible Harm** (P3, P4, P15, P17). Some focused on PII potentially leading to real-world losses. P17 explained: *"My phone number could be used for harassment calls, my home address could allow criminals to find me offline, and my ID number might be exploited for fraudulent loans. I cannot accept any of these being used as training data."*
- **Non-Consensual Data Collection** (P7, P11, P16, P18). A group of participants placed greater emphasis on unauthorized use of

PII. P11 remarked: *"No matter what type of PII it is, if I knowingly chose to make it public on the Internet, then I can accept it being used. Conversely, if information I never consented to share online is exposed, I cannot accept it—even if it is the most trivial PII."*

- **Risks of Profiling and Re-identification** (P19). One participant (P19) highlighted risks of aggregated user profiling. P19 explained: *"If multiple pieces of PII can be integrated to construct a complete profile of me, enabling others to predict my preferences and behaviors, then PII should never be acceptable as training data."*

Finding II-4: Participants deemed unacceptable PII for LLM training primarily as high-risk information, including directly identifiable, sensitive data or PII that could cause tangible harm, and as data lacking user autonomy, such as information collected without consent or not previously leaked. These views emphasize users' core concerns with risk severity and control.

5.5 User Responses to LLM Privacy Leakage (RQ3)

5.5.1 Continued Usage and Platform Migration (IQ16). Despite awareness of potential PII leakage, participants demonstrated strong continued usage intentions toward LLMs. No participant indicated plans to discontinue use, with many citing substantial efficiency and productivity benefits as overriding concerns. As P4 explained: *"The greatest advantage of LLMs is that they save me an enormous amount of time and improve my work efficiency. Compared to this advantage, privacy drawbacks are not enough for me to abandon LLMs."*

Similarly, most participants showed minimal motivation to switch to alternative LLM platforms. Only two participants (P9, P10) reported experimenting with alternative models for privacy protection. P9 considered transitioning to foreign-developed LLMs: *"I believe foreigners have less interest in my personal information, and even if they are interested, it would be harder for them to cause me harm."* P10 contemplated using models with reduced computational capacity: *"Perhaps with lower computing power, LLMs cannot train on such large volumes of personal data, and thus would not cause as much privacy leakage."*

5.5.2 Behavioral Adaptation in LLM Usage (IQ17). Six participants expressed willingness to modify their interaction patterns with LLMs to enhance privacy protection (P5, P6, P8, P11, P15, P16). These adaptations primarily involved avoiding disclosure of personal information during interactions. P16 stated: *"I will no longer provide LLMs with any of my real personal information."*

The majority of participants, however, remained unwilling to alter their usage practices. Some justified this stance by referencing pre-existing protective measures. P3 remarked: *"My existing usage practices already prevent privacy leakage quite well, so I see no need to change."* Others expressed skepticism regarding the effectiveness of behavioral changes, noting that their data may already reside within training datasets. P7 explained: *"My privacy has likely already been included in LLM training data, so changing my habits now cannot prevent LLMs from leaking it to others."*

Finding II-5: Despite recognizing PII leakage risks, participants maintained *strong continued usage intentions* due to perceived efficiency benefits. Most demonstrated limited willingness to modify their LLM interaction patterns, with some citing existing protective measures and others expressing *privacy cynicism* regarding the fundamental preventability of leakage.

5.6 User-Proposed Privacy Mitigation Strategies

Finally, we asked participants for their suggestions for mitigating privacy leakage in LLMs (IQ18). The recommendations, ranked by frequency of mention, are summarized below:

- **Training Data Transparency and Governance** (P2, P4, P5, P9, P10, P13, P16, P17, P20). The most frequently mentioned recommendations focused on enhanced transparency and regulatory oversight of training datasets. Many emphasized the need for LLM providers to openly disclose data acquisition methods. P10 stated: *"For data obtained via web crawling, they should clearly disclose which datasets were used and which websites were covered. For licensed datasets from companies, they should also specify the sources and broadly describe what the data includes."* Several participants advocated for governmental involvement in dataset regulation. P16 noted: *"If business interests or intellectual property prevent full disclosure, at the very least there should be a regulatory authority that can review the datasets and test whether they contain personal privacy or infringing content."*
- **Enhanced User Control Mechanisms** (P2, P4, P6, P8, P11, P13, P15, P18). Many participants suggested implementing enhanced user control mechanisms. P4 proposed: *"LLMs could add a function allowing users to designate a specific conversation or uploaded file as excluded from training."* P11 recommended incorporating consent revocation features: *"LLMs should support a withdrawal option for privacy consent. If I once agreed to provide conversations for training but later change my mind, there should be a feature that completely erases those conversations from memory."*
- **Data Curation and Anonymization** (P3, P5, P7, P9, P14, P15, P16). Improved pre-training data filtering practices were emphasized by several participants. P14 suggested: *"LLMs should either avoid collecting personal data during crawling or screen it out before training."* P15 proposed data anonymization approaches: *"Training data containing PII could be retained, but identifiers should be replaced with anonymized placeholders before being used."*
- **Real-Time Input and Output Safeguards** (P1, P2, P5, P8, P17, P18, P20). Multiple participants recommended adding defensive mechanisms at interaction points. P1 advocated for input monitoring systems: *"If an LLM detects that an upload or text input contains personal information, it should remind the user and allow them to decide whether to continue."* P5 suggested output filtering mechanisms: *"Before releasing an output, the system could scan whether the response contains personal data and intercept or process it further before displaying."*
- **Strengthened Government Supervision** (P3, P10, P12, P14, P19, P20). Strengthened governmental supervision was frequently proposed. P12 recommended systematic auditing: *"Government authorities could periodically review LLMs to check whether personal information has been misused or leaked at scale."* Others emphasized public education initiatives. P20 argued: *"Large-scale*

Table 7: LLM Objective Knowledge Assessment Items and Descriptive Statistics (n = 204)

Statements (Correct Answer in Brackets)	Mean	SD	Correct n	Correct Rate (%)
LLM training data mainly comes from the Internet. [T]	0.93	0.25	190	93.1
PII has been completely removed during the training of LLMs. [F]	0.88	0.32	180	88.2
LLM training datasets are fully open and publicly available. [F]	0.64	0.48	130	63.7
LLMs will never disclose PII contained in their training data. [F]	0.78	0.41	160	78.4
PII provided by users during conversations with LLMs may be further used for model training. [T]	0.91	0.28	186	91.2
The content users input into LLMs will never be disclosed to others. [F]	0.74	0.44	150	73.5
Publicly available information on the Internet does not contain PII. [F]	0.67	0.47	136	66.7
LLM training datasets do not include licensed proprietary data. [F]	0.61	0.49	125	61.3
Total Objective Literacy Score (0–8)	**6.16**	**2.05**	–	–

education initiatives are needed to raise awareness that LLMs may leak personal information. Many people are still unaware of this."

- **Restrictions on Web Scraping Practices** (P1, P7, P19). Several technically knowledgeable participants highlighted web scraping restrictions as essential. P7 suggested: *"Unauthorized web crawlers should be banned at the root, and data obtained this way should never be used commercially."* P19 recommended implementing crawling grace periods: *"There could be a rule prohibiting crawlers from accessing new websites for several months, giving site developers time to identify and fix potential privacy leaks themselves."*

Finding II-6: Users' suggested privacy safeguards converge on three areas: *greater transparency and governance of training data*, *enhanced user control over personal information*, and *stronger technical protections in data processing and interaction*. These measures aim to balance usability with privacy protection, thereby reducing user concerns and supporting the long-term use of LLMs.

6 Quantitative Survey Findings

In this section, we report participants' factual knowledge regarding LLM privacy risks (§6.1). We then analyze their main usage purposes for LLMs and identify the categories of PII considered most unacceptable for leakage (§6.2). Finally, we examine their threat perceptions and behavioral intentions regarding continued LLM usage despite these concerns (§6.3). Survey questions are denoted with the prefix "SQ" corresponding to items in Appendix C.

6.1 Factual Knowledge of LLM Privacy Risks (RQ2)

To assess users' factual understanding of privacy risks associated with LLMs, we adapted eight knowledge items (SQ8) from prior objective literacy measures [22, 95, 102], refined according to unawareness themes identified during interviews (see §5.2). Each item featured a single technically correct answer, with correct responses scored as 1 and incorrect responses as 0. Descriptive statistics for each item, including mean scores, standard deviations (SD), and correctness rates, are presented in Table 7.

Participants demonstrated relatively high objective literacy concerning LLM privacy risks, achieving a mean total score of 6.16 out of 8 (SD=2.05). This indicates substantial user awareness of fundamental aspects of LLM operations, such as the Internet-sourced nature of training data and the potential reuse of user-provided PII in model training, with accuracy rates exceeding 88–93%. However, knowledge regarding LLM training data sources—including whether datasets are publicly disclosed or incorporate licensed proprietary data—showed markedly lower accuracy, ranging from 61% to 64%. Understanding of LLM-related and internet-related privacy leakage, such as the potential disclosure of PII from training data or user inputs, or the presence of PII in publicly available web data, also proved less consistent, with accuracy rates of 67%-78%.

Finding III-1: Survey results indicate that although users possess a solid grasp of basic LLM functionalities, their knowledge regarding training data composition remains limited. Users tend to overestimate the privacy-preserving capabilities of LLMs and retain significant misconceptions about risks associated with personal information available on public web sources. These observations align closely with the unawareness dimensions identified in our qualitative interviews.

6.2 LLM Usage Purposes and PII Leakage Concerns (RQ3)

We examined both the primary use cases for LLMs and the categories of PII that users find most unacceptable to leak. Two multiple-choice questions were designed for this purpose; resulting statistics are summarized in Table 8.

Regarding usage purposes (SQ9), the predominant applications were *information retrieval* (93.6%), *daily work assistance* (81.9%), and *writing support* (77.9%). These were followed by *creative content generation* (46.1%), *learning or exam preparation* (42.6%), and *translation between Chinese and English* (42.2%). In contrast, *coding and debugging* (26.0%) and *casual chatting or entertainment* (31.4%) were less frequently reported, suggesting that users primarily employ LLMs for practical tasks rather than recreational or highly specialized functions.

With respect to PII leakage nonacceptance (SQ10), users expressed greatest concern toward highly sensitive categories, including *national ID or passport numbers* (96.1%), *bank or payment account information* (90.2%), and *home addresses* (83.3%). Additionally, *personal photos/audio/video* (70.6%) and *phone numbers* (64.7%) were frequently identified, consistent with the *Sensitive Personal Information* and *PII Leading to Physical or Financial Harm* categories discussed in §5.4. A considerable proportion of respondents also viewed *social media content* (62.3%) and *IP addresses/device fingerprints* (66.2%) as risky. By comparison, *email addresses* (26.0%) and *educational background* (26.0%) were perceived as less sensitive.

Table 8: LLM Usage Purposes and PII Leakage Nonacceptance (n = 204)

LLM Usage Purposes		PII Leakage Nonacceptance	
Usage Purpose	**Count (%)**	**PII Category**	**Count (%)**
Information retrieval / knowledge search	191 (93.6%)	Name	88 (43.1%)
Writing assistance	159 (77.9%)	ID / Passport number	196 (96.1%)
Daily chatting / entertainment	64 (31.4%)	Phone number	132 (64.7%)
Chinese ↔ Foreign language translation	86 (42.2%)	Email address	53 (26.0%)
Code writing and debugging	53 (26.0%)	Home address / Location information	170 (83.3%)
Daily work assistance	167 (81.9%)	Bank card / Payment account information	184 (90.2%)
Learning or exam preparation support	87 (42.6%)	Personal photos / Videos / Audio	144 (70.6%)
Creative content generation	94 (46.1%)	Medical records / Health data	85 (41.7%)
Other	0 (0.0%)	Educational background / School information	53 (26.0%)
		Employer / Job position	80 (39.2%)
		Online account / credentials	0 (0.0%)
		Social media posted content	127 (62.3%)
		IP address / Device information / Browser fingerprint	135 (66.2%)

Table 9: Construct Reliability, Descriptive Statistics, and Distribution of LLM Privacy Perceptions (n = 204)

Constructs	Cronbach's α	Mean	SD	Median	Low Level (≤3)	Medium Level (3–5)	High Level (≥5)
Perceived Vulnerability (SQ11)	0.878	4.77	1.21	5.00	29 (14.2%)	52 (25.5%)	123 (60.3%)
Perceived Severity (SQ12)	0.819	5.65	1.05	6.00	9 (4.4%)	29 (14.2%)	166 (81.4%)
Privacy Concern (SQ13)	0.908	5.02	1.17	5.30	19 (9.3%)	64 (31.4%)	121 (59.3%)
Trust (SQ14)	0.831	4.26	1.18	4.33	40 (19.6%)	109 (53.4%)	55 (27.0%)
Personal Interest (SQ15)	0.871	4.90	1.12	5.33	19 (9.3%)	76 (37.3%)	109 (53.4%)

Finding III-2: Consistent with interview findings on PII nonacceptance in §5.4, users report highest levels of concern and resistance toward the exposure of directly identifiable personal information and data related to physical or financial security during LLM interactions.

6.3 Privacy Risk Perceptions and Behavioral Intentions (RQ3)

6.3.1 Descriptive Analysis. Measurement results for user constructs in privacy-related contexts, including means, standard deviations, medians, and score distributions across different levels (*low level ≤3, medium level 3–5, high level ≥5*), are provided in Table 9. We examined five dimensions: *perceived vulnerability*, *perceived severity*, *trust*, *privacy concern*, and *personal interest* (SQ11-SQ15). All constructs demonstrated high internal consistency, with Cronbach's α values exceeding 0.80, indicating strong scale reliability.

Users generally perceived substantial privacy threats during LLM interactions. Specifically, *perceived severity* averaged 5.65 (on a 7-point scale), with 81.4% of users classified in the high-level group (scores≥5), reflecting widespread belief in serious consequences from privacy leakage. Similarly, *perceived vulnerability* averaged 4.77, with over 60% of users in the high-level group, indicating broad awareness of personal exposure likelihood. *Privacy concern* reflected strong threat perceptions, with an average score of 5.02 and nearly 60% of users falling into the high-concern group, underscoring substantial worries about leakage risks. In contrast, *trust* received a moderate mean score of 4.26, with a dispersed distribution: 53.4% of users were at medium trust levels, suggesting unstable or contested confidence in LLM providers' privacy protections.

At the behavioral level, *personal interest* averaged 4.90, falling in the medium-to-high range, with 53.4% of participants in the high-level group. This indicates that despite significant cognitive awareness of privacy threats, users maintain relatively high willingness to continue using LLMs, corroborating interview findings that practical benefits often outweigh privacy concerns.

Finding III-3: Consistent with the interview findings, the survey results indicate that users perceive LLM-related PII leakage as both highly likely and severe, while simultaneously reporting low levels of trust in LLM systems. Yet, despite these pronounced concerns, they continue to express strong intentions to use LLM services, underscoring the paradox between perceived risks and sustained adoption.

6.3.2 Correlation Analysis with Objective Privacy Literacy. To investigate how users' objective understanding of LLM-related privacy issues associates with their perceptions and behavioral intentions, we conducted correlation analyses between objective privacy literacy (detailed in §6.1) and the five key dimensions previously examined. As summarized in Table 10, objective literacy demonstrated significant positive correlations with perceived vulnerability ($r = .344, p < .001$), perceived severity ($r = .155, p = .030$), and overall privacy concern ($r = .202, p = .004$). These relationships suggest that users with higher privacy literacy possess greater awareness of potential privacy threats associated with LLMs. Additionally, we observed a significant negative correlation with trust in LLM providers ($r = -.281, p < .001$), indicating enhanced skepticism among more knowledgeable users.

Notably, however, objective literacy showed no statistically significant association with personal usage interest ($r = -.032, p = .651$). This pattern reveals a distinct *privacy literacy–usage paradox*: although more informed users recognize heightened risks and express diminished trust, this awareness does not correspond to reduced behavioral intention to utilize LLM services.

Table 10: Correlations between Objective Privacy Literacy and Other Variables

Variable	Correlation (r)	Significance (p)
Perceived Vulnerability	.344	<.001
Perceived Severity	.155	.030
Privacy Concern	.202	.004
Trust	-.281	<.001
Personal Interest	.032	.651

Finding III-4: Our correlation analysis demonstrates that users with higher privacy literacy exhibit increased awareness of LLM-related privacy risks and decreased trust in LLM systems. However, enhanced literacy shows no meaningful association with reduced LLM usage intention. This finding underscores a clear privacy paradox in which users' cognitive recognition of risks fails to manifest in altered behavioral patterns toward LLM adoption.

7 Discussion

7.1 Current Status of PII Leakage in Mainstream LLMs

7.1.1 Effectiveness: Pervasive Success of PII Extraction Attacks. Our evaluation reveals substantial privacy leakage risks in mainstream LLMs under both targeted and non-targeted extraction attacks. In targeted extraction, email addresses—as highly formatted and frequent PII types—were extracted with high success rates (average ASR: 78.3%), indicating easy memorization and reproduction by LLMs. Phone#, due to sparse distribution and heterogeneous formats [81], showed lower but still significant leakage (average ASR: 20.3%), confirming vulnerability even for less-structured PII.

In non-targeted extraction, large-scale harvesting was feasible across professions. Doctors and journalists faced highest risks (average ASRs: 38.1% and 42.8%, respectively), with some LLM models exceeding 60–70% leakage. Although accountants and lawyers showed relatively lower averages (34.6% and 33.7%), their vulnerability remained non-negligible. Overall, 6,919 authentic PII instances were harvested with an average ASR of 34.6%, meaning one in three attempts successfully retrieved valid personal records. These results demonstrate that PII extraction is a practical, large-scale threat in real-world settings, indicating systemic rather than isolated risks.

7.1.2 Variability: Cross-Model Differences in Leakage Susceptibility. Significant cross-model variability emerged in privacy leakage susceptibility. In targeted attacks, most models performed similarly for email extraction, but phone number success rates varied widely—*e.g.*, *gpt-5* and *chatgpt-4o* exceeded 39%, while *qwen3-235b-a22b* and *glm-4.5* dropped to 5–7%. Non-targeted extraction showed greater divergence: *chatgpt-4o* and *deepseek-v3* reached 59.3% and 49.8% ASR, whereas *qwen3-235b-a22b* and *hunyuan-turbos* remained at 15.6% and 20.2%.

This variability may stem from differences in training corpus composition, formatted vs. unformatted data prevalence, model architecture, data distribution, and safety alignment strategies [118]. Language coverage also likely played a role: since ground-truth datasets and few-shot examples were English-based, models trained primarily on non-English corpora may underperform in English extraction [121, 126]. Indeed, Chinese-developed models like *qwen3-235b-a22b* and *hunyuan-turbos* [112] scored below average, suggesting linguistic composition significantly influences PII extraction performance. These findings underscore the need for model-specific defense strategies rather than one-size-fits-all approaches.

7.1.3 Severity: Challenges in Training Data Protection. Our experiments confirm that PII from public LLM training datasets can be directly extracted from model outputs with high success rates and broad applicability. This not only validates LLMs' memorization and regurgitation of training data but also highlights the inadequacy of current defenses.

Our verification process is necessarily limited to PII that remains publicly accessible through conventional web search, but a substantial proportion of even such PII originates from user-unauthorized platforms, including illicit data markets and gray-market data brokers that systematically disclose personal information without consent while remaining indexed by major search engines [24, 28]. LLMs inadvertently compound these privacy violations by aggregating, reorganizing, and amplifying dispersed personal data across their training corpora. More concerningly, as demonstrated by Cheng *et al.* [28], LLMs can memorize and subsequently leak PII that has been deliberately removed from active web presence but persists in historical archives such as the Internet Archive and Common Crawl snapshots. Most critically, the PII extraction methodologies we employ are fundamentally generalizable and model-agnostic in their design. This characteristic implies that their effectiveness extends beyond publicly available training datasets. If publicly sourced PII proves extractable with such efficiency, then proprietary licensed corpora and user interaction data incorporated into model training likely face comparable memorization and leakage vulnerabilities without stronger safeguards. This broader implication raises serious concerns regarding the long-term security and privacy sustainability of LLM deployment across real-world applications.

7.2 User Literacy, Attitudes, and Perceptions of LLM Privacy Leakage

Both interview and survey results reveal a complex pattern of user perceptions characterized by limited knowledge, heightened concern, and contradictory attitudes.

7.2.1 Knowledge Gaps. Users exhibited significant knowledge gaps regarding LLM training data composition and extraction techniques, especially concerning non-public sources and interaction data reuse. Many assumed training relied solely on public or licensed datasets, unaware that everyday LLM interactions could be incorporated. Users also underestimated public Internet PII risks, overlooking sensitivity arising from model aggregation. Survey results confirmed stronger awareness of basic LLM functions but sharp accuracy drops (61–66%) on questions about dataset disclosure and proprietary data sources. Overall, users possess only surface-level literacy, with broad misconceptions leading to reliance on intuition rather than systematic judgment.

7.2.2 Divergent Attitudes and Emotional Responses. Attitudes toward LLM privacy leakage were predominantly negative. Many

expressed fear and anxiety over undisclosed datasets, aggregation capabilities, and potential misuse, describing a "loss of control" or feeling like a "transparent person." These reactions highlight perceived vulnerability and resistance to technological opacity. Conversely, some participants adopted neutral or cynical stances, arguing that privacy is already widely compromised and LLM leakage would not fundamentally change the situation. Such privacy cynicism may reduce resistance but also undermine public support for stronger protections [30, 101].

7.2.3 Balancing Utility and Privacy Concerns. A key finding is the tension between utility and concern. Despite widespread apprehension, many users evaluated training data collection positively, citing irreplaceable benefits of LLMs in efficiency, knowledge access, and productivity. They engaged in mental risk-benefit calculus consistent with the *Privacy Calculus* model [34, 36]. Although users recognize severe privacy threats, they maintain strong usage intentions and show little inclination to modify behavioral practices. This misalignment between awareness and behavior underscores the real-world dilemma of LLM privacy leakage.

7.3 User Responses to Privacy Risks and Future Implications

7.3.1 User Risk Acceptance and Behavioral Inertia. Evidence from interviews and surveys indicates users generally lack robust strategies to address LLM privacy leakage. Most participants showed neither willingness to discontinue use nor motivation to switch platforms. They emphasized that LLMs' efficiency, productivity, and convenience outweigh privacy risks, reflecting "continued adoption under known risks." Moreover, most were reluctant to modify practices—some believing existing routines sufficed, others exhibiting privacy cynicism that individual changes would not yield meaningful protection. This mindset reflects powerlessness against powerful technical systems and highlights limitations of user-driven measures.

This observed pattern of "continued adoption under known risks" constitutes a concrete manifestation of the privacy paradox in the LLM context. While some scholarship challenges the conceptual validity of the privacy paradox—arguing that researchers mistakenly interpret context-specific disclosures as indicators of low overall privacy concern [60, 107]—our findings indicate that the attitude–behavior discrepancy in LLM usage may arise from several intertwined mechanisms. The inherent opacity of LLM training data collection and processing, together with the technical complexity of these systems, may limit users' understanding of how privacy breaches actually occur, as shown in §5.2 and §6.2. This lack of clarity makes long-term risks difficult to evaluate and can render the risk component of the privacy calculus cognitively inaccessible, which helps explain the absence of correlation between objective privacy literacy and usage intention in our survey result in §6.3.2. In addition, LLM PII extractions are a relatively recent line of research that has emerged only within the past year and have not yet been applied in real-world exploitation, further reducing users' ability to perceive concrete threats. At the same time, the exceptional functional capabilities of LLMs contribute to structural dependencies and a limited availability of viable alternatives, which may encourage continued usage despite privacy concerns, as we observed in §5.5.1. Finally, the widespread nature of data breaches across the broader digital ecosystem can foster privacy cynicism, in which users come to doubt that individual protective measures would meaningfully improve their privacy posture [49]. Taken together, these dynamics suggest that the privacy paradox in this context cannot be attributed simply to users "not caring about privacy," nor dismissed as a purely conceptual artifact; instead, it reflects a complex socio-technical phenomenon shaped by systemic opacity, structural lock-in, limited risk visibility, cognitive burdens, and broader feelings of cynicism, which renders the paradox empirically meaningful in practical contexts [45].

7.3.2 Rejected PII Categories: User-Defined Privacy Boundaries. Participants also articulated clear boundaries regarding unacceptable PII types for LLM training. They rejected high-risk information (directly identifiable or harm-inducing data) and data lacking user autonomy (collected without consent or not previously exposed). These criteria underscore two central concerns: severity of potential harm and degree of control over personal data. Such insights provide valuable guidance for future safeguards and emphasize the importance of transparency and consent mechanisms in privacy governance [61, 72].

7.3.3 User-Proposed Privacy Safeguards: Aspirations and Practical Constraints. Participants' proposed privacy safeguards directly reflected their underlying concerns, emphasizing three primary categories: enhanced transparency and regulatory oversight of training data sources, improved user control mechanisms including data exclusion and revocation capabilities, and strengthened real-time safeguards during model interactions. Collectively, these suggestions indicate that users value not only technical effectiveness of defenses but also institutional and structural accountability within the LLM ecosystem [38, 41, 72]. Participants envisioned a comprehensive, multi-layered protection framework integrating corporate responsibility, technical reinforcement, governmental regulation, and public awareness initiatives [61, 89, 125]. However, these user-identified safeguard categories face distinct challenges regarding effectiveness, technical feasibility, and practical implementation:

(1) **Transparency and Regulatory Oversight Enhancement:** While this approach garners broad ethical and governance support, its practical implementation presents significant hurdles. Comprehensive disclosure of training corpora may compromise proprietary information and potentially introduce new attack vectors, such as targeted data poisoning attacks on internet-crawled sources [4]. Furthermore, effective regulatory enforcement depends on legislative advancement, international coordination, and platform compliance—factors characterized by inherently slow developmental trajectories.
(2) **User Control Mechanisms Augmentation:** Several LLM providers have begun implementing features enabling users to opt out of personal data utilization for training purposes [6, 85]. Nevertheless, enabling retroactive data revocation presents substantial technical obstacles. Complete model retraining entails prohibitive computational costs, while existing machine unlearning and model editing techniques frequently contend with persistent memory effects, rendering comprehensive data removal unreliable in practice [8, 54, 77].

(3) **Real-time Safeguard Reinforcement:** Comparative analysis suggests real-time protective mechanisms offer greater technical implementability. Many contemporary LLMs incorporate external content moderation systems designed to flag or block user inputs containing prohibited content categories such as hate speech, violence, or dangerous activities [86]. However, these systems typically prioritize broad content categorization rather than specialized PII detection. Additionally, such moderation frameworks remain vulnerable to circumvention through evolving jailbreak techniques. Current research initiatives focusing on real-time PII detection within LLMs demonstrate promising directions but often introduce trade-offs in model utility degradation and increased inference latency [31, 42, 104, 110].

In summary, while users conceptualize an ideal privacy protection paradigm based on socio-technical co-governance, current technological capabilities and platform implementation practices remain insufficient to fully realize these expectations. This implementation gap not only mirrors users' fundamental concerns regarding safety and autonomy in digital environments but also underscores the pressing need to develop privacy-preserving mechanisms that simultaneously achieve technical robustness, practical feasibility, and operational transparency.

7.4 Ethics Considerations

The data extracted from public LLMs in our experiments may contain real-world private information. To adhere to ethical principles [97], we implemented standard practices: reporting only aggregate results, permanently deleting all extracted and analyzed PII, and presenting anonymized results to participants before prompt deletion. Our university IRB approved this approach. While our research inevitably highlights methods to induce sensitive content generation, we believe raising awareness of these vulnerabilities is crucial for enabling targeted defenses. We have responsibly disclosed findings to relevant LLM vendors.

7.5 Limitations

This study has several limitations. First, our PII extraction evaluation framework inherits limitations from the incorporated extraction methods, including potential false negatives due to the inability to verify the authenticity of extracted PII originating from non-public datasets. Additionally, the ASR for some LLMs may be underestimated due to insufficient parameter tuning for individual models. Second, the participant pool's composition—primarily drawn from academic settings—may limit the generalizability of our qualitative findings. Individuals from this demographic typically exhibit stronger verbal articulation and analytical abilities than the general population, potentially resulting in responses that appear more polished and conceptually refined. This sampling approach may underrepresent perspectives from other demographic groups, thereby constraining the broader applicability of our insights. Moreover, our user study predominantly involved Chinese participants, limiting generalizability to other regions, as privacy perceptions are shaped by cultural and institutional contexts.

8 Conclusion

This study reveals the current state of privacy leakage in mainstream LLMs and how users perceive and respond to this issue. Our evaluation shows that PII can still be efficiently extracted from public training datasets (78.3% for emails, 20.3% for phone numbers), with large-scale non-targeted attacks reaching 34.6% and some models exceeding 50%, highlighting the limits of existing safeguards. Interviews and surveys further indicate that users lack knowledge of training data practices and underestimate leakage risks, yet remain highly concerned about directly identifiable and harmful PII. Despite low trust and privacy cynicism, users continue to adopt LLMs while calling for greater transparency, control, and institutional safeguards. These findings underscore both the technical and user dimensions of LLM privacy leakage and provide implications for designing trustworthy models that balance efficiency with privacy.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62272410).

References

[1] [n. d.]. Serper. Serper. https://serper.dev/.
[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv:2303.08774* (2023).
[3] Atilla Akkus, Mingjie Li, Junjie Chu, Michael Backes, Yang Zhang, and Sinem Sav. 2024. Generated data with fake privacy: Hidden dangers of fine-tuning large language models on generated data. *arXiv:2409.11423* (2024).
[4] Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. 2025. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine* 31, 2 (2025), 618–626.
[5] Mutahar Ali, Arjun Arunasalam, and Habiba Farrukh. 2025. Understanding Users' Security and Privacy Concerns and Attitudes Towards Conversational AI Platforms. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 298–316.
[6] Anthropic. 2024. Anthropic Privacy Policy. https://www.anthropic.com/legal/privacy.
[7] Anthropic. 2025. Claude Opus 4 and Sonnet 4 Model Card. https://www.anthropic.com/transparency. Accessed: 2025-09-08.
[8] Tomer Ashuach, Martin Tutek, and Yonatan Belinkov. 2024. REVS: Unlearning Sensitive Information in Language Models via Rank Editing in the Vocabulary Space. *arXiv:2406.09325* (2024).
[9] Yang Bai, Ge Pei, Jindong Gu, Yong Yang, and Xingjun Ma. 2024. Special characters attack: Toward scalable training data extraction from large language models. *arXiv:2405.05990* (2024).
[10] Evan Bailyn. 2025. ChatGPT Usage Statistics. First Page Sage Blog. https://firstpagesage.com/seo-blog/chatgpt-usage-statistics
[11] Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2024. Emergent and predictable memorization in large language models. *NeurIPS* 36 (2024).
[12] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*. PMLR, 2397–2430.
[13] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745* (2022).
[14] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *If you use this software, please cite it using these metadata* 58, 2 (2021).
[15] Nadine Bol, Tobias Dienlin, Sanne Kruikemeier, Marijn Sax, Sophie C Boerman, Joanna Strycharz, Natali Helberger, and Claes H De Vreese. 2018. Understanding the effects of personalization as a privacy calculus: Analyzing self-disclosure across health, news, and commerce contexts. *Journal of Computer-Mediated Communication* 23, 6 (2018), 370–388.

[16] Jaydeep Borkar. 2023. What can we learn from Data Leakage and Unlearning for Law? *arXiv:2307.10476* (2023).

[17] Jaydeep Borkar, Matthew Jagielski, Katherine Lee, Niloofar Mireshghallah, David A Smith, and Christopher A Choquette-Choo. 2025. Privacy Ripple Effects from Adding or Removing Personal Information in Language Model Training. *arXiv preprint arXiv:2502.15680* (2025).

[18] Amber Bouman. 2025. *Meta AI was leaking chatbot prompts and answers to unauthorized users.* https://www.tomsguide.com/computing/online-security/meta-ai-was-leaking-chatbot-prompts-and-answers-to-unauthorized-users Published: 17 July 2025; Accessed: 2025-09-10.

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

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

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

[22] Jay P Carlson, William O Bearden, and David M Hardesty. 2007. Influences on what consumers know and what they think they know regarding marketer pricing tactics. *Psychology & Marketing* 24, 2 (2007), 117–142.

[23] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How people use chatgpt.* Technical Report. National Bureau of Economic Research.

[24] Guangxuan Chen, Qiang Liu, Guangxiao Chen, and Anan Huang. 2025. Exploring illicit personal information trading behind telecom fraud in China. *Humanities and Social Sciences Communications* 12, 1 (2025), 1–11.

[25] Ruizhe Chen, Tianxiang Hu, Yang Feng, and Zuozhu Liu. 2024. Learnable Privacy Neurons Localization in Language Models. *arXiv:2405.10989* (2024).

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

[27] Shuai Cheng, Zhao Li, Shu Meng, Mengxia Ren, Haitao Xu, Shuai Hao, Chuan Yue, and Fan Zhang. 2025. Understanding PII Leakage in Large Language Models: A Systematic Survey. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI-25)*. International Joint Conferences on Artificial Intelligence Organization.

[28] Shuai Cheng, Shu Meng, Haitao Xu, Haoran Zhang, Shuai Hao, Chuan Yue, Wenrui Ma, Meng Han, Fan Zhang, and Zhao Li. 2025. Effective {PII} Extraction from {LLMs} through Augmented {Few-Shot} Learning. In *34th USENIX Security Symposium (USENIX Security 25)*. 8155–8173.

[29] Christy M Cheung and Matthew K Lee. 2001. Trust in internet shopping: instrument development and validation through classical and modern approaches. *Journal of Global Information Management (JGIM)* 9, 3 (2001), 23–35.

[30] Hanbyul Choi, Jonghwa Park, and Yoonhyuk Jung. 2018. The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior* 81 (2018), 42–51.

[31] Chun Jie Chong, Chenxi Hou, Zhihao Yao, and Seyed Mohammadjavad Seyed Talebi. 2024. Casper: Prompt Sanitization for Protecting User Privacy in Web-Based Large Language Models. *arXiv:2408.07004* (2024).

[32] Common Crawl. 2025. Common Crawl. https://commoncrawl.org.

[33] Credamo. 2025. Credamo – Intelligent Research Platform. https://www.credamo.cc/. An online platform offering services such as questionnaire design, sample recruitment, and data analytics.

[34] Mary J Culnan and Pamela K Armstrong. 1999. Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization science* 10, 1 (1999), 104–115.

[35] Ajinkya Deshmukh, Saumya Banthia, and Anantha Sharma. 2023. Life of PII–A PII Obfuscation Transformer. *arXiv:2305.09550* (2023).

[36] Tamara Dinev and Paul Hart. 2006. An extended privacy calculus model for e-commerce transactions. *Information systems research* 17, 1 (2006), 61–80.

[37] Fabio Duarte. 2025. Number of ChatGPT Users (July 2025). https://explodingtopics.com/blog/chatgpt-users.

[38] Md Meftahul Ferdaus, Mahdi Abdelguerfi, Elias Ioup, Kendall N Niles, Ken Pathak, and Steven Sloan. 2024. Towards trustworthy ai: A review of ethical and robust large language models. *arXiv preprint arXiv:2407.13934* (2024).

[39] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions.* john wiley & sons.

[40] Donna L Floyd, Steven Prentice-Dunn, and Ronald W Rogers. 2000. A meta-analysis of research on protection motivation theory. *Journal of applied social psychology* 30, 2 (2000), 407–429.

[41] Vincent Freiberger, Arthur Fleig, and Erik Buchmann. 2025. " You don't need a university degree to comprehend data protection this way": LLM-Powered Interactive Privacy Policy Assessment. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[42] Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. IncogniText: Privacy-enhancing Conditional Text Anonymization via LLM-based Private Attribute Randomization. *arXiv:2407.02956* (2024).

[43] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027* (2020).

[44] Leo Gao, Sid Black, Stella Biderman, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. OpenWebText2: An Improved Open-source WebText Corpus. https://github.com/EleutherAI/openwebtext2. EleutherAI Project.

[45] Nina Gerber, Paul Gerber, and Melanie Volkamer. 2018. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & security* 77 (2018), 226–261.

[46] Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus: An Open-source Replication of the WebText Dataset. https://skylion007.github.io/OpenWebTextCorpus/. Accessed: 2025-09-08.

[47] Ece Gumusel, Kyrie Zhixuan Zhou, and Madelyn Rose Sanfilippo. 2024. User privacy harms and risks in conversational ai: A proposed framework. *arXiv preprint arXiv:2402.09716* (2024).

[48] Anil Gurung, Xin Luo, and Qinyu Liao. 2009. Consumer motivations in taking action against spyware: An empirical investigation. *Information Management & Computer Security* 17, 3 (2009), 276–289.

[49] Christian Pieter Hoffmann, Christoph Lutz, and Giulia Ranzini. 2016. Privacy cynicism: A new approach to the privacy paradox. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 10, 4 (2016).

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

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

[52] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[53] Princely Ifinedo. 2012. Understanding information systems security policy compliance: An integration of the theory of planned behavior and the protection motivation theory. *Computers & Security* 31, 1 (2012), 83–95.

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

[55] Seyoung Jin, Heewon Baek, Uichin Lee, and Hyoungshick Kim. 2025. I Was Told to Install the Antivirus App, but I'm Not Sure I Need It: Understanding Smartphone Antivirus Software Adoption and User Perceptions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[56] Allen C Johnston and Merrill Warkentin. 2010. Fear appeals and information security behaviors: An empirical study. *MIS quarterly* (2010), 549–566.

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

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

[59] Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*. Springer, 217–226.

[60] Spyros Kokolakis. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & security* 64 (2017), 122–134.

[61] Jabari Kwesi, Jiaxun Cao, Riya Manchanda, and Pardis Emami-Naeini. 2025. 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)*. 6007–6024.

[62] Matthew KO Lee and Efraim Turban. 2001. A trust model for consumer internet shopping. *International Journal of electronic commerce* 6, 1 (2001), 75–91.

[63] Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. 2013. What matters to users? factors that affect users' willingness to share information with online advertisers. In *Proceedings of the ninth symposium on usable privacy and security*. 1–12.

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

[65] Huigang Liang, Yajiong Lucky Xue, et al. 2010. Understanding security behaviors in personal computer usage: A threat avoidance perspective. *Journal of the association for information systems* 11, 7 (2010), 1.

[66] Fangyu Lin, Laura Brandimarte, Sue Brown, and Hsinchun Chen. 2024. Examining the Effect of Personalized PII Exposure Alerts on Individuals' Privacy Protection Motivation. (2024).

[67] Zhihuang Liu, Ling Hu, Tongqing Zhou, Yonghao Tang, and Zhiping Cai. 2025. Prevalence Overshadows Concerns? Understanding Chinese Users' Privacy Awareness and Expectations Towards LLM-Based Healthcare Consultation. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2716–2734.

[68] LMArena Team. 2025. Leaderboard Changelog: notable updates to model leaderboards. https://news.lmarena.ai/leaderboard-changelog/.

[69] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664* (2021).

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

[71] Lumivero. 2025. NVivo: Leading Qualitative Data Analysis Software. Lumivero product page. https://lumivero.com/products/nvivo/

[72] Rongjun Ma, Caterina Maidhof, Juan Carlos Carrillo, Janne Lindqvist, and Jose Such. 2025. Privacy perceptions of custom gpts by users and creators. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[73] Ying Ma, Shiquan Zhang, Dongju Yang, Zhanna Sarsenbayeva, Jarrod Knibbe, and Jorge Goncalves. 2025. Raising Awareness of Location Information Vulnerabilities in Social Media Photos using LLMs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–14.

[74] Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R Trippas, J Shane Culpepper, and Alistair Moffat. 2020. CC-News-En: A large English news corpus. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 3077–3084.

[75] Miguel Malheiros, Charlene Jennett, Snehalee Patel, Sacha Brostoff, and Martina Angela Sasse. 2012. Too close for comfort: A study of the effectiveness and acceptability of rich-media personalized advertising. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 579–588.

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

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

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

[79] Sarah Milne, Paschal Sheeran, and Sheina Orbell. 2000. Prediction and intervention in health-related behavior: A meta-analytic review of protection motivation theory. *Journal of applied social psychology* 30, 1 (2000), 106–143.

[80] Ariffud Muhammad. 2025. LLM Statistics 2025: Comprehensive Insights Into Market Trends and Integration. Hostinger Tutorials. https://www.hostinger.com/tutorials/llm-statistics/

[81] Krishna Kanth Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. PII-Compass: Guiding LLM training data extraction prompts towards the target PII via grounding. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*. 63–73.

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

[83] Nate Nelson. 2024. *Hundreds of LLM Servers Expose Corporate, Health & Other Online Data.* https://www.darkreading.com/application-security/hundreds-of-llm-servers-expose-corporate-health-and-other-online-data Published: 28 August 2024; Accessed: 2025-09-10.

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

[85] OpenAI. 2023. OpenAI Privacy Policy. https://openai.com/policies/privacy-policy.

[86] OpenAI. 2025. Content Moderation – OpenAI Platform. https://platform.openai.com/docs/guides/moderation.

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

[88] Ashwinee Panda, Christopher A. Choquette-Choo, Zhengming Zhang, Yaoqing Yang, and Prateek Mittal. 2024. Teach llms to phish: Stealing private information from language models. In *ICLR*. https://openreview.net/forum?id=qo21ZlfNu6

[89] Emmanouil Papagiannidis, Patrick Mikalef, and Kieran Conboy. 2025. Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems* 34, 2 (2025), 101885.

[90] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116* (2023).

[91] Jeanne Pigassou and Rayan Ben Taleb. 2025. *Leaking Minds: How Your Data Could Slip Through AI Chatbots.* https://www.riskinsight-wavestone.com/en/2025/05/leaking-minds-how-your-data-could-slip-through-ai-chatbots/ Accessed: 2025-09-10.

[92] Jiantao Qiu, Haijun Lv, Zhenjiang Jin, Rui Wang, Wenchang Ning, Jia Yu, ChaoBin Zhang, Zhenxiang Li, Pei Chu, Yuan Qu, et al. 2024. Wanjuan-cc: A safe and high-quality open-sourced english webtext dataset. *arXiv preprint arXiv:2402.19282* (2024).

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

[94] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[95] Puthankurissi S Raju, Subhash C Lonial, and W Glynn Mangold. 1995. Differential effects of subjective knowledge, objective knowledge, and usage experience on decision making: An exploratory investigation. *Journal of consumer psychology* 4, 2 (1995), 153–180.

[96] Md Rafi Ur Rashid, Jing Liu, Toshiaki Koike-Akino, Ye Wang, and Shagufta Mehnaz. 2025. Forget to Flourish: Leveraging Machine-Unlearning on Pretrained Language Models for Privacy Leakage. *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (2025).

[97] Caitlin M Rivers and Bryan L Lewis. 2014. Ethical research standards in a world of big data. *F1000Research* 3 (2014), 38.

[98] Ronald W Rogers. 1975. A protection motivation theory of fear appeals and attitude change1. *The journal of psychology* 91, 1 (1975), 93–114.

[99] Ronald W Rogers. 1983. Cognitive and physiological processes in fear appeals and attitude change: A revised theory of protection motivation. *Social psychology: A source book* (1983), 153–176.

[100] Sruly Rosenblat, Tim O'Reilly, and Ilan Strauss. 2025. Beyond Public Access in LLM Pre-Training Data. *arXiv preprint arXiv:2505.00020* (2025).

[101] Wilmar B Schaufeli. 1996. Maslach burnout inventory-general survey (MBI-GS). *Maslach burnout inventory manual* (1996).

[102] Claire M Segijn, Eunah Kim, Asma Sifaoui, and Sophie C Boerman. 2023. When you realize that big brother is watching: How informing consumers affects synced advertising effectiveness. *Journal of Marketing Communications* 29, 4 (2023), 317–338.

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

[104] Li Siyan, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. 2024. PAPILLON: PrivAcy Preservation from Internet-based and Local Language MOdel ENsembles. *arXiv preprint arXiv:2410.17127* (2024).

[105] H Jeff Smith, Sandra J Milberg, and Sandra J Burke. 1996. Information privacy: Measuring individuals' concerns about organizational practices. *MIS quarterly* (1996), 167–196.

[106] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159* (2024).

[107] Daniel J Solove. 2021. The myth of the privacy paradox. *Geo. Wash. L. Rev.* 89 (2021), 1.

[108] Qiurong Song, Yanlai Wu, Rie Helene Hernandez, Yao Li, Yubo Kou, and Xinning Gui. 2025. Understanding Users' Perception of Personally Identifiable Information. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–24.

[109] Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. 208–220.

[110] Xiongtao Sun, Gan Liu, Zhipeng He, Hui Li, and Xiaoguang Li. 2024. DePrompt: Desensitization and Evaluation of Personal Identifiable Information in Large Language Model Prompts. *arXiv:2408.08930* (2024).

[111] Anthropic Research Team. 2025. *Anthropic Economic Index: September 2025 Report – Uneven geographic and enterprise AI adoption.* Research Report. Anthropic. https://www.anthropic.com/research/anthropic-economic-index-september-2025-report

[112] Tencent Hunyuan Team, Ao Liu, Botong Zhou, Can Xu, Chayse Zhou, ChenChen Zhang, Chengcheng Xu, Chenhao Wang, Decheng Wu, Dengpeng Wu, et al. 2025. Hunyuan-turbos: Advancing large language models through mamba-transformer synergy and adaptive chain-of-thought. *arXiv preprint arXiv:2505.15431* (2025).

[113] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
[114] Unknown. 2025. *The Grandma Exploit Explained.* https://jailbreakai.substack.com/p/the-grandma-exploit-explained-prompt Accessed: 2025-09-10; Substack article, author not specified.
[115] Davide Venditti, Elena Sofia Ruzzetti, Giancarlo A Xompero, Cristina Giannone, Andrea Favalli, Raniero Romagnoli, and Fabio Massimo Zanzotto. 2024. Enhancing Data Privacy in Large Language Models through Private Association Editing. *arXiv:2406.18221* (2024).
[116] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models.. In *NeurIPS*.
[117] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.
[118] Shang Wang, Tianqing Zhu, Bo Liu, Ming Ding, Xu Guo, Dayong Ye, Wanlei Zhou, and Philip S Yu. 2024. Unique security and privacy threats of large language model: A comprehensive survey. *arXiv preprint arXiv:2406.07973* (2024).
[119] Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. 2024. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems* 37 (2024), 116462–116492.
[120] Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. Depn: Detecting and editing privacy neurons in pretrained language models. *arXiv:2310.20138* (2023).
[121] Zhong Yao, Liantan Duan, Shuo Xu, Lingyi Chi, and Dongfang Sheng. 2025. Performance of Large Language Models in the Non-English Context: Qualitative Study of Models Trained on Different Languages in Chinese Medical Examinations. *JMIR Medical Informatics* 13, 1 (2025), e69485.
[122] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.
[123] Zhexin Zhang, Jiaxin Wen, and Minlie Huang. 2023. Ethicist: Targeted training data extraction through loss smoothed soft prompting and calibrated confidence estimation. In *ACL*. 12674–12687.
[124] Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2024. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for Navigating Privacy Trade-offs in LLM-Based Conversational Agent. *arXiv:2410.11876* (2024).
[125] John JianJun Zhu, Ling Tuo, Yanfen You, Qiang Fei, and Matthew Thomson. 2024. A preemptive and curative solution to mitigate data breaches: Corporate social responsibility as a double layer of protection. *Journal of Marketing Research* 61, 4 (2024), 778–801.
[126] Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948* (2023).
[127] Yixin Zou, Khue Le, Peter Mayer, Alessandro Acquisti, Adam J Aviv, and Florian Schaub. 2024. Encouraging users to change breached passwords using the protection motivation theory. *ACM Transactions on Computer-Human Interaction* 31, 5 (2024), 1–45.

A Interview Questionnaire

This section presents the semi-structured interview guide used in our study on user perceptions and responses to PII leakage in LLMs. The interview comprises three parts.

A.1 Part 1: Demographic and LLM Usage Information

- IQ1: What is your age range?
  - 18–27
  - 28–37
  - 38–47
  - 48–57
  - 58+
- IQ2: What is your gender?
  - Male
  - Female
  - Other
  - Prefer not to say
- IQ3: What is your current employment status?
  - Student
  - Employed
  - Other
  - Prefer not to say
- IQ4: Do you have a computer science-related job or educational background?
  - CS-related
  - Non-CS
- IQ5: Which LLM(s) do you commonly use? *[open-ended question]*
- IQ6: What is your LLM usage frequency?
  - Daily
  - Weekly
  - Monthly
  - Rarely

A.2 Part 2: Experiences and Initial Views of LLM Privacy Leakage

- IQ7: Do you believe LLMs are capable of leaking personally identifiable information (PII)? Why or why not?
  - Do you think it is possible to deliberately use LLMs to obtain others' personal information?
- IQ8: Have you personally experienced situations where LLMs appeared to leak PII (either yours or others')?
- IQ9: Would you voluntarily provide your PII to LLMs during interaction? Why or why not?

Technical Information Session. Before proceeding to the next section, participants were introduced to background knowledge (mainly selected technical details and experimental illustrations, with attack demonstrations omitted).

(a) LLM Training Data

LLMs rely on large training datasets, which can be divided into two main types:

- **Public sources**: scraped from the open Internet, including:
  - Web pages (blogs, forums, articles)
  - Books and academic publications
  - Open-source code repositories (*e.g.*, GitHub)
  - Social platform discussions (*e.g.*, Reddit)

  Notable examples include *Enron*, *The Pile*, *Common Crawl*, and *OpenWebText*. These datasets frequently contain PII—such as names, email addresses, phone numbers, or sensitive details inadvertently shared in forums or issue trackers. In some instances, public datasets also include leaked or misconfigured data, where users never intended their private information to be exposed (*e.g.*, due to website misconfigurations resulting in unintended disclosures). [*Related examples omitted*]
- **Non-public sources**: used by commercial models, including licensed content from companies and user interaction data such as inputs, outputs, account registration info,

logs, and cookies. Privacy policies of systems like ChatGPT and Claude state that user inputs may, in some cases, be reused for further training, unless disabled by enterprise agreements. [*Related examples omitted*]

Companies such as OpenAI, Anthropic, or Google do not disclose full details of their training datasets, but research has shown that traces of public datasets appear in their outputs, indicating persistent PII risks.

(b) PII Extraction Attacks

Researchers have developed various *attack strategies* to extract PII from LLMs [a selection of representative methods with demonstrations is summarized below; detailed descriptions are omitted]:

- **Jailbreaking Prompts**: Carefully engineered prompts constructed to circumvent model safeguards and elicit unauthorized disclosures.
- **Chain-of-Thought (CoT) Prompting**: Multi-step reasoning prompts that incrementally lead a model toward the divulgence of confidential information.
- **Few-Shot Prompting**: Supplying a limited number of structured examples to prompt the bulk generation of PII adhering to analogous formats.
- **Fine-Tuning**: The process of updating models with new datasets, which can intensify memorization and induce leakage from both the new data and the original training corpus.

Researchers evaluate these techniques using two types of attacks:

- **Targeted Extraction**: The retrieval of data pertaining to a specific individual (*e.g.*, a particular individual's telephone number).
- **Non-Targeted Extraction**: The bulk harvesting of PII from numerous individuals (*e.g.*, compiling the email addresses of medical practitioners).

(c) Our Evaluation of LLMs

To ascertain the severity of PII leakage, we employed these extraction techniques on ten prominent commercial LLMs (including GPT-5, Claude, Gemini, among others). Both targeted and non-targeted extraction were evaluated using public ground-truth datasets. [Related demonstrations omitted]

- **Targeted Extraction**: We attained notably high success rates in recovering email addresses (average 78.3%), whereas the extraction of phone numbers was less successful yet remained considerable (20.3%). This indicates that standardized identifiers such as email addresses are extensively memorized and susceptible to retrieval. [*Performance comparison table omitted*]
- **Non-Targeted Extraction**: We successfully extracted thousands of authentic name-email-phone triples across various professional domains, including medical, legal, and journalistic fields, with an average success rate of approximately 34.6%. Specific models exhibited leakage exceeding 60% for queries within particular professions. [*Extraction results figure omitted*]

(d) Summary

- The training data for LLMs is predominantly acquired through Internet crawls and may incorporate non-public sources; the precise composition of these datasets is not disclosed.
- Such datasets invariably contain PII, and user-submitted inputs may in some instances be repurposed for training.
- Consequently, LLMs are capable of memorizing and potentially exposing sensitive information, including names, email addresses, and telephone numbers.
- Empirical investigations confirm that the effective and efficient extraction of such PII from mainstream LLMs is presently feasible.
- Collectively, these findings demonstrate that all these vectors are susceptible to exploitation via technical attacks.

A.3 Part 3: User Literacy, Attitudes, and Behavioral Intentions Regarding PII Leakage

Privacy Literacy and Awareness Gaps.

- IQ10: Before this interview, which aspects of LLM background knowledge in Part 2 were you unaware of?

Perspectives on LLM Privacy Leakage.

- IQ11: What are your overall views on the current state of LLM privacy leakage?
- IQ12: How do you evaluate the sources and collection practices of LLM training data?
  - Do you think such data collection should be conducted?
- IQ13: How do you evaluate the extent of your own PII leakage on the Internet?
- IQ14: Are you concerned that your leaked PII may be collected and exposed by LLMs? Why?

Acceptance Dimensions of PII Usage in LLM Training.

- IQ15: Which types of PII do you find acceptable or unacceptable for use in LLM training?

User Responses.

- IQ16: Would you stop using or switch to alternative LLMs due to privacy concerns? Why or why not?
- IQ17: Would you modify your personal LLM usage practices? Why or why not?

User Suggestions.

- IQ18: What suggestions would you give to LLM developers or platforms to mitigate privacy leakage?

B Constructs and Measurement Items in Our Survey

Table 11 presents the constructs, their measurement items, and sources, as utilized in our survey.

C Survey Questionnaire

The following presents the complete questionnaire on users' literacy and perceptions of PII Leakage of LLM.

C.1 Part 1: Demographic Information

SQ1: What is your gender? *[Single choice]*

- Male
- Female

Table 11: Items for Each Construct in Our Survey

Construct	Items	Ref
Objective Privacy Literacy	SQ8: Please evaluate whether the following statements are correct. *[Matrix, Single choice for each item: Correct / Incorrect / I don't know]* – LLM training data mainly comes from the Internet. – PII has been completely removed during the training of LLMs. – LLM training datasets are fully open and publicly available. – LLMs will never disclose PII contained in their training data. – PII provided by users during conversations with LLMs may be further used for model training. – The content users input into LLMs will never be disclosed to others. – Publicly available information on the Internet does not contain PII. – LLM training datasets do not include licensed proprietary data.	[22, 95, 102]
Perceived Vulnerability	SQ11: How likely do you think the following events could happen to you? *[Matrix, 7-point Likert: Very Unlikely – Very Likely]* – My PII may be leaked by LLM to other users. – I may become a target of privacy attacks when using LLMs. – PII that I provide in LLM interactions may be accessed by others.	[48, 55, 99, 127]
Perceived Severity	SQ12: How severe would the impact be if the following events occurred? *[Matrix, 7-point Likert: Not Severe at All – Very Severe]* – My PII is used as LLM training data. – My PII is leaked due to LLM outputs. – Criminals misuse leaked PII (*e.g.*, account theft, spam). – Economic losses caused by LLM-related privacy leakage.	[48, 55, 99, 127]
Privacy Concern	SQ13: How concerned are you about the following possibilities when using LLMs? *[Matrix, 7-point Likert: No Concern at All – Extremely High Concern]* – I am concerned that the information I provide to LLMs could be misused. – I am concerned that others may obtain my private information through LLMs. – I am concerned about submitting personal information to LLMs, because of how the model or others might use it. – I am concerned about submitting personal information to LLMs, because it could be used in a way I did not foresee.	[34, 36, 105]
Trust	SQ14: Please indicate your agreement with the following statements. *[Matrix, 7-point Likert: Strongly Disagree – Strongly Agree]* – LLMs are safe environments for providing personal information. – LLMs are reliable environments for important tasks and work. – LLMs handle user-submitted PII professionally and responsibly.	[29, 36, 62]
Personal Interest	SQ15: Please indicate your agreement with the following statements. *[Matrix, 7-point Likert: Strongly Disagree – Strongly Agree]* – I find that personal interest in the information that I want to obtain from LLMs overrides my concerns of possible risk or vulnerability that I may have regarding PII leakage. – The greater my interest or needs to obtain a certain information or service from LLMs, the more I tend to suppress my privacy concerns. – In general, my need to obtain certain information or services from LLMs is greater than my concern about privacy risks.	[15, 36]

  - Other
  - Prefer not to disclose

SQ2: What is your age? *[Single choice]*
  - 18–27
  - 28–37
  - 38–47
  - 48–57
  - 58 or older
  - Prefer not to disclose

SQ3: What is your highest level of education completed? *[Single choice]*
  - Middle school or below
  - High school / Vocational school
  - Associate degree
  - Bachelor's degree
  - Master's degree
  - Doctoral degree
  - Prefer not to disclose

SQ4: What is your current occupation? *[Single choice]*
  - Student
  - Freelancer
  - Self-employed
  - Government employee
  - Professional (*e.g.*, doctor, lawyer, journalist, teacher)
  - Corporate manager
  - General office staff
  - Agricultural / Fishing / Forestry worker
  - Retired
  - Prefer not to disclose

SQ5: Do you have a background in computer science, telecommunications, or artificial intelligence? *[Single choice]*
  - Yes
  - No
  - Prefer not to disclose

SQ6: What is your total annual income this year? *[Single choice]*
  - Less than 100,000 RMB
  - 100,000–200,000 RMB
  - 200,000–400,000 RMB

- 400,000–600,000 RMB
- 600,000–800,000 RMB
- 800,000–1,000,000 RMB
- More than 1,000,000 RMB
- Prefer not to disclose

SQ7: How frequently do you use large language models? *[Single choice]*
- Daily
- Weekly
- Monthly
- Occasionally
- Tried only once
- Prefer not to disclose

C.2 Part 2: Perceptions of LLM Privacy Leakage

Before answering the following questions, please carefully read the description below:

Large Language Models (LLMs) denote widely-used models including ChatGPT, Claude, DeepSeek, Gemini, and Qwen, which are capable of performing tasks such as dialogue, content generation, programming assistance, and information retrieval.

Personally Identifiable Information (PII) encompasses private data that may be utilized to identify a specific individual. Examples include full name, government-issued identification numbers, phone number, email address, residential address, account credentials, location data, photographs, or medical history. Unauthorized disclosure of such information could lead to misuse including identity theft, telecommunications fraud, or other unlawful activities.

Research has shown that during user interactions, LLMs may produce outputs containing PII, thereby introducing risks associated with privacy leakage.

SQ8: Please evaluate whether the following statements are correct. *[Matrix, Single choice for each item: Correct / Incorrect / I don't know]*
- LLM training data mainly comes from the Internet.
- PII has been completely removed during the training of LLMs.
- LLM training datasets are fully open and publicly available.
- LLMs will never disclose PII contained in their training data.
- PII provided by users during conversations with LLMs may be further used for model training.
- The content users input into LLMs will never be disclosed to others.
- Publicly available information on the Internet does not contain PII.
- LLM training datasets do not include licensed proprietary data.

SQ9: For what purposes do you typically use large language models? *[Multiple choice]*
- Information retrieval / knowledge search
- Writing assistance (*e.g.*, papers, emails, copywriting)
- Daily chatting / entertainment
- Chinese ↔ Foreign language translation
- Code writing and debugging
- Daily work assistance (*e.g.*, data analysis, document generation)
- Learning or exam preparation support
- Creative content generation (*e.g.*, stories, poetry, image prompts)
- Other: _____
- Prefer not to disclose

SQ10: Which types of PII do you perceive as most unacceptable if leaked by LLMs? *[Multiple choice]*
- Name
- ID / Passport number
- Phone number
- Email address
- Home address / Location information
- Bank card / Payment account information
- Personal photos / Videos / Audio
- Medical records / Health data
- Educational background / School information
- Employer / Job position
- Online account / credentials
- Social media posted content
- IP address / Device information / Browser fingerprint

SQ11: How likely do you think the following events could happen to you? *[Matrix, 7-point Likert: Very Unlikely – Very Likely]*
- My PII may be leaked by LLM to other users.
- I may become a target of privacy attacks when using LLMs .
- PII that I provide in LLM interactions may be accessed by others.

SQ12: How severe would the impact be if the following events occurred? *[Matrix, 7-point Likert: Not Severe at All – Very Severe]*
- My PII is used as LLM training data.
- My PII is leaked due to LLM outputs.
- Criminals misuse leaked PII (*e.g.*, account theft, spam).
- Economic losses caused by LLM-related privacy leakage.

SQ13: How concerned are you about the following possibilities when using LLMs? *[Matrix, 7-point Likert: No Concern at All – Extremely High Concern]*
- I am concerned that the information I provide to LLMs could be misused.
- I am concerned that others may obtain my private information through LLMs.
- I am concerned about submitting personal information to LLMs, because of how the model or others might use it.
- I am concerned about submitting personal information to LLMs, because it could be used in a way I did not foresee.

SQ14: Please indicate your agreement with the following statements. *[Matrix, 7-point Likert: Strongly Disagree – Strongly Agree]*
- LLMs are safe environments for providing personal information.
- LLMs are reliable environments for important tasks and work.
- LLMs handle user-submitted PII professionally and responsibly.

Table 12: Demographics in the Online Survey

Gender	N (%)
Male	120 (58.8%)
Female	94 (41.2%)
Other	0 (0.0%)
Prefer not to disclose	0 (0.0%)
Age	N (%)
18–27	108 (52.9%)
28–37	56 (27.4%)
38–47	24 (11.8%)
48–57	15 (7.4%)
58 or older	1 (0.5%)
Prefer not to disclose	0 (0.0%)
Education Level	N (%)
Middle school or below	1 (0.5%)
High school / Vocational school	8 (3.9%)
Associate degree	15 (7.4%)
Bachelor's degree	147 (72.1%)
Master's degree	32 (15.7%)
Doctoral degree	1 (0.5%)
Prefer not to disclose	0 (0.0%)
Occupation	N (%)
Student	54 (26.5%)
Freelancer	8 (3.9%)
Self-employed	3 (1.5%)
Government employee	11 (5.4%)
Professional	17 (8.3%)
Corporate manager	39 (19.1%)
General office staff	69 (33.8%)
Agricultural / Fishing / Forestry worker	2 (1.0%)
Retired	1 (0.5%)
Prefer not to disclose	0 (0.0%)
CS/AI Background	N (%)
Yes	125 (61.3%)
No	79 (38.7%)
Prefer not to disclose	0 (0.0%)
Annual Income	N (%)
Less than 100,000 RMB	84 (41.2%)
100,000–200,000 RMB	80 (39.2%)
200,000–400,000 RMB	30 (14.7%)
400,000–600,000 RMB	5 (2.5%)
600,000–800,000 RMB	2 (1.0%)
800,000–1,000,000 RMB	0 (0.0%)
More than 1,000,000 RMB	0 (0.0%)
Prefer not to disclose	3 (1.5%)
LLM Usage Frequency	N (%)
Daily	101 (49.5%)
Weekly	83 (40.7%)
Monthly	12 (5.9%)
Occasionally	7 (3.4%)
Tried only once	1 (0.5%)
Prefer not to disclose	0 (0.0%)

SQ15: Please indicate your agreement with the following statements. *[Matrix, 7-point Likert: Strongly Disagree – Strongly Agree]*
- I find that personal interest in the information that I want to obtain from LLMs overrides my concerns of possible risk or vulnerability that I may have regarding PII leakage.
- The greater my interest or needs to obtain a certain information or service from LLMs, the more I tend to suppress my privacy concerns.
- In general, my need to obtain certain information or services from LLMs is greater than my concern about privacy risks.

D Demographics of Online Survey Participants

Table 12 summarizes the demographic characteristics of the 204 valid respondents in our online survey. The survey encompassed gender, age, education, occupation, background in computer science or artificial intelligence, annual income, and LLM usage frequency, aligning with the options listed in the questionnaire (SQ1–SQ7).

E Interview Codebook

Table 13 presents the finalized interview codebook derived from our qualitative study. The codebook was constructed through thematic analysis of 20 semi-structured interviews and organizes participants' responses into six major categories: *(i)* experiences and initial perceptions of LLM privacy leakage, *(ii)* gaps in awareness regarding LLM-related privacy literacy, *(iii)* attitudes toward LLM privacy leakage and training data collection practices, *(iv)* dimensions of nonacceptance toward PII usage in LLM training, *(v)* user reactions to LLM privacy leakage, and *(vi)* user-proposed measures for mitigating privacy risks associated with LLMs. Each category is further subdivided into sub-codes, accompanied by illustrative examples and counts of participant responses.

Table 13: Codebook for Interview

No.	Codes	# Responses
A	**Experiences and Initial Views of LLM Privacy Leakage**	
A.1	Belief that LLMs can leak PII	15
A.1.1	Personal experiences suggesting leakage	5
A.1.2	Technical reasoning based on training data or mechanisms	7
A.1.3	Intuitive judgment	3
A.2	Reported experiences of potential leakage	6
A.3	Reluctance to provide PII during interaction	16
A.3.1	Avoidance or anonymization of PII before input	10
B	**Gaps in LLM Privacy Literacy**	
B.1	Limited Understanding of LLM Fundamentals	13
B.2	Limited Awareness of PII Extraction Attacks	5
B.3	Underestimation of Risks of PII in Public Website	8
B.4	Underestimation of PII Extraction Efficacy	7
C	**Perspectives on LLM Privacy Leakage**	
C.1.1	Negative perceptions of current LLM privacy leakage status	13
C.1.1.1	Fear of undisclosed datasets	5
C.1.1.2	Feelings of helplessness	4
C.1.1.3	Concerns about criminal misuse	6
C.1.2	Neutral or less negative perceptions of current LLM privacy leakage status	7
C.1.2.1	Belief that leakage is not unique to LLMs	4
C.1.2.2	Privacy cynicism	4
C.2.1	Positive views on training data collection practices	16
C.2.1.1	Legality or legitimacy of collection	6
C.2.1.2	Necessity for improving performance	12
C.2.2	Negative views on training data collection practices	11
C.2.2.1	Unauthorized scraping	4
C.2.2.2	Ethical concerns	2
C.2.2.3	Hidden PII in licensed datasets	6
C.3.1	Severe exposure perception of personal PII exposure on the Internet	15
C.3.1.1	Social engineering risks	7
C.3.1.2	Cross-platform linkages	5
C.3.1.3	Firsthand experience of leakage	4
C.3.2	Moderate exposure perception of personal PII exposure on the Internet	5
C.3.2.1	Due to proactive privacy measures	5
C.4.1	High concern about LLM-related PII leakage	11
C.4.1.1	Aggregation capacity of LLMs	7
C.4.1.2	Sense of exposure	4
C.4.2	Moderate or low concern about LLM-related PII leakage	9
C.4.2.1	Considered non-valuable target	4
C.4.2.2	Belief that traditional methods are more effective	5
D	**Unacceptable PII Categories for LLM Training**	
D.1	Directly identifying PII (*e.g.*, ID, phone, address)	10
D.2	Sensitive personal information (*e.g.*, health, finance)	6
D.3	PII entailing tangible harm	4
D.4	Non-consensual data collection	4
D.5	Perceived novelty of exposure	2
D.6	Risks of Profiling and Re-identification	1
E	**User Responses to LLM Privacy Leakage**	
E.1.1	Platform migration	2
E.1.1.1	Switch to foreign-developed LLMs	1
E.1.1.2	Switch to LLMs with reduced computational capacity	1
E.1.2	Continued Usage without platform migration	18
E.1.2.1	Driven by efficiency and productivity	18
E.2.1	Willingness to modify practices	6
E.2.1.1	Avoid providing PII to LLMs	6
E.2.2	Reluctance to modify practices	14
E.2.2.1	Already practicing cautious behavior	8
E.2.2.2	Skepticism about the effectiveness of usage modifications	6
F	**User-Proposed Privacy Mitigation Strategies**	
F.1	Training data transparency and governance	9
F.2	Enhanced user control mechanisms	8
F.3	Data curation and anonymization	7
F.4	Real-Time input and output safeguards	7
F.5	Strengthened government supervision	6
F.6	Restricting on web scraping practices	3