

POSITION: PRIVACY IS NOT JUST MEMORIZATION!

Niloofer Mireshghallah^{1*} Tianshi Li^{2*}

¹Carnegie Mellon University ²Northeastern University

niloofer@cmu.edu tia.li@northeastern.edu

ABSTRACT

The discourse on privacy risks in Large Language Models (LLMs) has disproportionately focused on verbatim memorization of training data, while a constellation of more immediate and scalable privacy threats remain underexplored. *This position paper argues that the privacy landscape of LLM systems extends far beyond training data extraction, encompassing risks from data collection practices, inference-time context leakage, autonomous agent capabilities, and the democratization of surveillance through deep inference attacks.* We present a comprehensive taxonomy of privacy risks across the LLM lifecycle—from data collection through deployment—and demonstrate through case studies how current privacy frameworks fail to address these multifaceted threats. Through a longitudinal analysis of 1,322 AI/ML privacy papers published at leading conferences over the past decade (2016–2025), we reveal that while memorization receives outsized attention in technical research, the most pressing privacy harms lie elsewhere, where current technical approaches offer little traction and viable paths forward remain unclear. We call for a fundamental shift in how the research community approaches LLM privacy, moving beyond the narrow focus of current technical solutions and embracing interdisciplinary approaches that address the sociotechnical nature of these emerging threats.

1 INTRODUCTION

Large Language Models are fundamentally data-driven systems, trained on vast corpora scraped from the web (Brown et al., 2020), user interactions (Ouyang et al., 2022), and increasingly, real-time retrieval systems (Lewis et al., 2020). While privacy concerns have rightfully emerged as these models consume unprecedented amounts of personal data, the research community’s response has been disproportionately narrow—fixating almost exclusively on verbatim memorization and training data extraction (Carlini et al., 2021; 2023). In this paper, we advance a critical position: *Privacy in LLM systems is not just about memorization. It encompasses how providers extract consent through deceptive interfaces, how autonomous agents exfiltrate user data without regard for privacy norms, how systems aggregate scattered information to reveal intimate details and provide answers to secondary questions used for password recovery, and how models can transform innocuous public data into targeted surveillance or stalking capabilities.*

We systematically categorize the privacy landscape by first identifying **three types of data** flowing through LLM ecosystems: *(i) user interaction data* encompassing prompts, feedback, and conversation histories; *(ii) system-retrieved data* from RAG pipelines, APIs, and real-time sources; and *(iii) publicly available data* including web corpora with embedded credentials and personal information (Section 2). These data types interact to create **five distinct categories of privacy incidents** (Table 1): beyond *training data leakage via regurgitation*, we identify *direct chat leakage* through provider breaches and deceptive policies, *indirect context leakage* via autonomous agents and prompt injection, *indirect attribute inference* where LLMs deduce sensitive information from innocuous inputs, and *direct attribute aggregation* that weaponizes dispersed online information (Section 3.1–3.5). Each incident type presents unique threats—from agents exfiltrating database contents through compromised RAG systems to LLMs inferring precise locations from seemingly anonymous photos—fundamentally transforming these systems from passive data stores into active, privacy-violating inference engines. *The harms extend beyond privacy into security domains*, as

*Equal Contribution, order decided by coin toss

Table 1: Taxonomy of personal data incidents in LLMs: We divide the incidents into five categories, over three different data types:  User interactions (§2.2.1),  Retrieved documents (§2.2.2), and  Publicly available data (§2.2.3). Victim is the person or entity whose data is revealed and data viewer is the entity that gains access to this revealed data, maliciously or by accident.

Section	Incident Type	Target Data	Victim	Data Viewer	Model Role
§3.1	Training Data Leakage via Regurgitation	 User interactions  Public data	 User  Bystander w/ public data	 Innocent user  Malicious user  Innocent bystander	Model as data-store
§3.2	Direct Chat Leakage via Uninformed Consent or Compromised Provider	 User interactions (Full transcript)	 User	 Innocent bystander  Legal proceedings  Malicious 3rd party	Model not directly involved
§3.3	Indirect Chat and Context Leakage via Input-Output Flow	 User interactions  Retrieved documents or data via API	 User	 Malicious 3rd party  Innocent bystander	Model as autonomous agent
§3.4	Indirect Attribute Inference	 Available data fed to LLM to infer age, location, etc.	 Bystander	 Malicious user	Model as inference engine
§3.5	Direct Attribute Aggregation	 Public data: finding exact attributes via deep research	 Bystander w/ public data	 Malicious user	Model as search engine

information aggregated by deep research agents can be exploited to answer seemingly innocuous questions—such as “What’s Alice’s pet cat’s name?” (see a real example in Figure 2)—which in turn can enable secondary attacks like password retrieval and account theft (Little et al., 2024).

Having identified five distinct categories of privacy incidents in LLM systems (Section 3)—from training data leakage to inference attacks and aggregation threats—a critical question emerges: does the research community’s focus align with these real-world privacy risks? To answer this, we conduct a systematic analysis of AI/ML privacy research published at leading conferences over the past decade (2016–2025). Our findings (Section 4) reveal a striking misalignment between research priorities and practical privacy threats. While 92% of papers focus on training data memorization and cryptographic protections against direct chat leakage, the remaining incident types—indirect attribute inference, agent-based context leakage, and direct attribute aggregation—collectively receive less than 8% of research attention suggesting disciplinary blind spots that leave critical vulnerabilities unaddressed.

This paper paves a path forward through technical interventions that work today (local data minimization, hybrid architectures, privacy-aligned post-training), sociotechnical approaches that empower users (contextual integrity frameworks, awareness tools, tradeoff visualization), and policy reforms that address power asymmetries. We demonstrate that privacy protection requires moving beyond the narrow lens of memorization to address deceptive consent, inference attacks, and the commodification of conversation. The privacy challenges are sociotechnical, not purely algorithmic—requiring collaboration between technologists, designers, policymakers, and affected communities. The rest of the paper is organized as follows:

§2 What Data is Affected?

- §2.1 Data Collection and Retention Policies
 - §2.1.1 What is Explicit Consent? The Default Opt-in Setting
 - §2.1.2 Do Users Really Have a Choice? Opt-out and Other Limitations
- §2.2 Different Types of Data in the LLM Ecosystem
 - §2.2.1 User Interaction Data
 - §2.2.2 System Retrieved Data
 - §2.2.3 Publicly Available Data

§3 How is the Data Being Exposed?

- §3.1 Training Data Leakage via Regurgitation
- §3.2 Direct Chat Leakage via Uninformed Consent or Compromised Provider
- §3.3 Indirect Chat and Context Leakage via Input-Output Flow
- §3.4 Privacy Under the Microscope: Indirect Attribute Inference
- §3.5 Privacy Through the Telescope: Direct Attribute Aggregation

§4 A Decade of AI/ML Privacy Research: Trends from Leading ML, NLP, and S&P Conferences

- §4.1 Corpus
- §4.2 Annotation Pipeline
- §4.3 Results

§5 Technical Solutions and Beyond: A Roadmap Forward

- §5.1 Technical Interventions
- §5.2 Sociotechnical Approaches
- §5.3 Policy and Governance

§6 Conclusion

2 WHAT DATA IS AFFECTED?

The scope of data at risk in LLM systems extends far beyond training corpora. To understand the privacy risks posed by LLMs beyond verbatim memorization and regurgitation of often publicly-available pretraining data (Carlini et al., 2021; 2023), we must expand our view from the model in isolation to the entire LLM ecosystem. This ecosystem encompasses data collection and curation, model training, deployment infrastructure, serving systems, and third-party API wrappers (Wang et al., 2025; Bommasani et al., 2021). Recent research has identified systematic vulnerabilities across this ecosystem, from poisoned RAG systems (Zou et al., 2025) to insecure third-party app stores (Hou et al., 2025). Each component in this ecosystem touches different types of data, creating compounding privacy vulnerabilities that current research has only begun to explore (Siyan et al., 2024; Liu et al., 2025b).

We categorize the data affected by the LLM ecosystem into three distinct types: *user interaction data*, *system-retrieved data*, and *publicly available data*. We define each category and analyze the unique privacy risks they present. These categories are not mutually exclusive—their risks compound when data flows between them, as we demonstrate through our incident taxonomy in Table 1. Understanding these data types and their interconnections is crucial for developing comprehensive privacy protections that address the full scope of LLM-related risks, moving beyond the narrow focus on training data extraction that has dominated prior work.

2.1 DATA COLLECTION AND RETENTION POLICIES

Before we discuss the different types of data impacted by the LLM ecosystem, we need to examine how LLM providers define consent, implement opt-out mechanisms, and retain user data in practice.

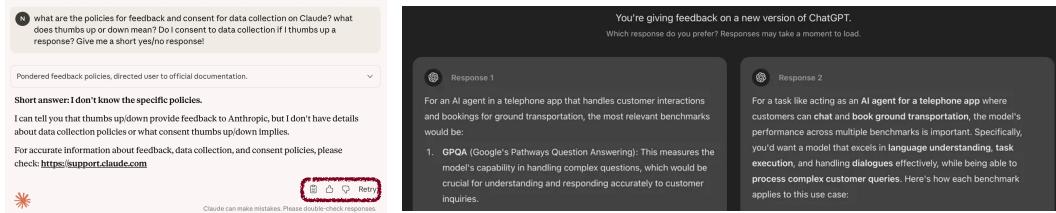


Figure 1: Examples of ‘automatic’ consent mechanisms deployed by Anthropic (giving a thumbs up or down on Claude responses opts the conversation into data collection, left) and OpenAI (selecting a response records the conversation in ChatGPT, right).

2.1.1 WHAT IS EXPLICIT CONSENT? THE DEFAULT OPT-IN SETTING

In the context of LLM services, explicit consent has been fundamentally redefined in ways that would be unrecognizable under traditional privacy frameworks. Let’s have a look at the different policies set by prominent frontier LLM providers:

- **Anthropic** employed what appeared to be the most restrictive approach on paper, stating they “will not use your Inputs or Outputs to train our generative models unless you’ve explicitly reported the materials to us.” This opt-in model applied equally to free and paid individual users, though enterprise customers received additional protections through Zero Data Retention agreements. **This restrictive approach was abandoned in September 2025 (the time of writing).** In a reversal from their privacy-first stance, Anthropic now requires users to explicitly opt-out by September 28, 2025, or their conversations and coding sessions will be used to train AI models. The new policy extends data retention from 30 days to 5 years for users who do not opt out Anthropic (2025c). This shift affects all consumer tiers (Claude Free, Pro, and Max), though enterprise customers under Commercial Terms maintain their privacy protections. **Note that automatic consent still occurs when users provide feedback through thumbs up/down mechanisms (see Figure 1), regardless of their opt-out status Anthropic (2025b).**
- **OpenAI** defines consent more broadly, with free users’ data used for training by default unless they actively opt out. Their policy states that “Content” includes “any data, files, or information you provide through our Services,” encompassing prompts, uploads, and all interactions OpenAI (2024a). Paid tiers like ChatGPT Plus follow the same default training usage as free users, while only enterprise customers receive automatic opt-out protections OpenAI (2024b).
- **Google Gemini** takes the most expansive approach, with “Gemini Apps Activity on by default if you are 18 or older,” automatically collecting “your chats, what you share with Gemini Apps (like files, images, screens), related product usage information, your feedback, and location info” (Google, 2025b). While Google One AI Premium subscribers receive some enhanced protections where “Google doesn’t use your prompts or responses to improve our products,” feedback and certain metadata remain subject to collection.
- **Grok/xAI** implements the most aggressive collection, with all X users automatically opted-in to data sharing for AI training. The November 2024 policy update expanded this to include sharing with “third-party collaborators” beyond just xAI (TechCrunch, 2024). This includes all public posts, interactions, voice inputs, and cross-platform integration data when using X credentials (xAI, 2025b).

In summary all major providers now operate on opt-out models that favor data collection. While the majority offer paid tiers that ostensibly provide enhanced privacy protections, even paid subscriptions contain hidden vulnerabilities where various mechanisms automatically opt you in or grant consent on your behalf. Let’s examine these practices.

Thumbs up or down? You just consented to 10 years of data retention, even as a paid user! Perhaps most concerning is how all providers exploit feedback mechanisms to bypass privacy protections. OpenAI explicitly states: “*Even if you’ve opted out of training, if you choose to provide*

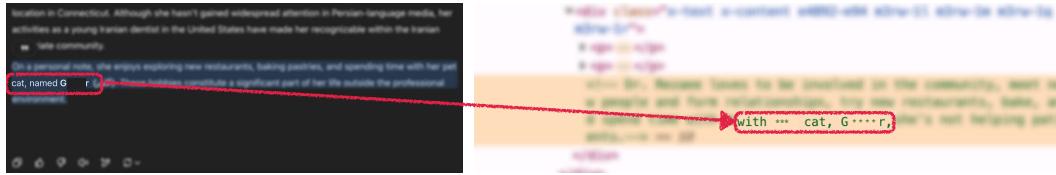


Figure 2: Example of a redacted query to ChatGPT’s deep research: It uncovers the name of an individual’s pet cat from a comment embedded in an HTML tag. This is particularly concerning, as such niche information is often used in password recovery, which could facilitate account theft and create security risks (Little et al., 2024).

feedback (for instance, by selecting thumbs up or thumbs down), the entire conversation associated with that feedback may be used to train our models” (OpenAI, 2024c). This creates a particularly deceptive practice where a simple evaluative gesture grants comprehensive training rights that override all other privacy settings. The “which is better” comparison interfaces employed by these services similarly trigger automatic data collection rights.

Anthropic retains feedback-related conversations “*in our secured back-end for up to 10 years*,” while Google keeps such data “for up to 3 years, disconnected from your Google Account” but *immune to user deletion requests* (Anthropic, 2025d; Google, 2025b). Even Grok’s feedback system creates permanent training data that *cannot be removed from models once processed* (xAI, 2025a). **Note that even after Anthropic’s September 2025 policy change requiring explicit opt-out for training, the feedback mechanism still triggers extended retention periods that override user preferences.**

Arbitrary security classifiers can mark to keep your data forever. All major providers maintain broad security exceptions that override deletion policies. OpenAI reserves the right to retain data for “legal or security reasons” with automated classifiers for abuse detection triggering minimum 30-day retention that *can extend indefinitely* (OpenAI, 2024d). Anthropic’s Constitutional AI classifiers trigger extended retention: “We retain inputs and outputs for up to 2 years and trust and safety classification scores for *up to 7 years* if you submit a prompt that is flagged by our trust and safety classifiers” (Anthropic, 2025d). These classifiers monitor for CBRN content, violence, and other policy violations with *opaque criteria*. Google maintains that “conversations that have been reviewed or annotated by human reviewers are not deleted when you delete your Gemini Apps activity,” with 3-year retention for flagged content (Google, 2025b). The company’s cross-service integration enables data sharing for “detecting, preventing, and responding to fraud, abuse, security risks, and technical issues” across all Google properties (Google, 2025).

Your conversations persist for years. Standard retention periods range from 30 days (Anthropic's default deletion for non-flagged conversations) **30 days only for Anthropic users who explicitly opt out (as of September 2025), to 5 years for those who don't opt out**, to 18 months (Google's default Gemini activity retention), with *feedback data retained for 3-10 years regardless of account deletion* (Anthropic, 2025d; Google, 2025b). More critically, *a federal court order since May 2025 requires OpenAI to preserve consumer ChatGPT and API customer data indefinitely indefinitely, even if they are deleted by the user*, affecting all consumer users (OpenAI, 2024; VentureBeat, 2025).

Training usage depends on both user tier and interaction type: Anthropic maintains its no-training policy without explicit consent; **Anthropic now uses consumer data by default unless users opt out (as of September 2025)**; OpenAI uses free/Plus user data unless opted out while protecting enterprise data by default; Google trains on all free user data while excluding paid user prompts but continuing to use their feedback; and Grok trains on all user data by default with *retroactive model training that cannot be reversed* (Anthropic, 2025b; OpenAI, 2024a; Google, 2025a; Silicon Republic, 2024).

2.1.2 DO USERS REALLY HAVE A CHOICE? OPT-OUT AND OTHER LIMITATIONS

While providers offer opt-out mechanisms, these systems are deliberately complex and often ineffective, creating barriers that discourage users from exercising privacy rights.

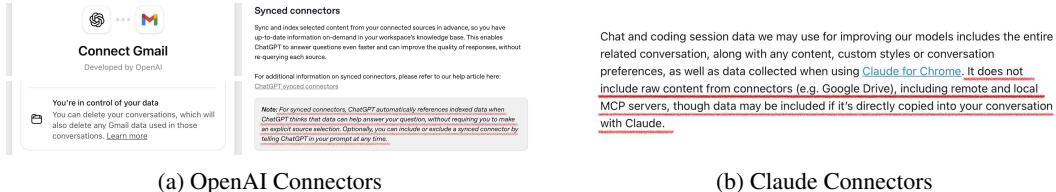


Figure 3: OpenAI and Claude both provide connectors for *automatic* integration of external user data, but the data is often scattered and requires manual deletion to be fully removed.

- **Anthropic** provides the most limited options with *no global opt-out for feedback data once submitted* and no retroactive control over previously shared conversations (Anthropic, 2025b). API customers face additional restrictions: “*For paid API customers, we do not support ad hoc deletion*” (Anthropic, 2025a).
- **OpenAI**’s multi-pathway opt-out system—accessible through ChatGPT Settings, privacy.openai.com, or Temporary Chat mode—appears comprehensive but contains critical gaps. The chat history used to be unnecessarily bundled with training controls (Zhang et al., 2024c) until it was separated into distinct controls in April 2024 (BGR, 2024). Moreover, the ongoing New York Times lawsuit has resulted in *a court order requiring OpenAI to preserve all user data indefinitely*, fundamentally undermining any opt-out preferences (OpenAI, 2024; Magai, 2025).
- **Google Gemini**’s opt-out mechanism has drawn particular criticism for its complexity. Users must navigate to myactivity.google.com, locate Gemini-specific controls among dozens of Google services, and even then face limitations: “Conversations that have been reviewed or annotated by human reviewers are not deleted when you delete your Gemini Apps activity” (Google, 2025b). The July 2025 update that *automatically enabled Gemini access to Phone, Messages, and WhatsApp data* “whether your Gemini Apps Activity is on or off” further eroded user control.
- **Grok/xAI** requires navigating to Settings → Privacy and Safety → Data sharing and personalization → Grok to disable training, but *data already processed cannot be removed from models*. The EU forced X to delete illegally processed data from May-August 2024, but this applied only to European users (RPC, 2024).

Beyond these contrived interfaces, additional concerning practices further undermine user control over their data.

“Temporary/vanish Chat” mode still keeps your data for 30 days. Features marketed as privacy-enhancing often provide minimal actual protection. ChatGPT’s “Temporary Chat” mode *still retains conversations for up to 30 days for “safety purposes”* (OpenAI, 2024e). Google’s “incognito” mode for Gemini similarly maintains retention for safety monitoring, while Grok’s “private” conversations remain visible to xAI for moderation purposes (Google, 2025b; xAI, 2025b).

Deletion is an option ... only on paper and not in practice! Despite offering deletion options, practical limitations severely restrict user control. *OpenAI’s consumer users cannot actually delete their data due to the federal court order*, with the standard 30-day deletion timeline overridden by indefinite legal hold (VentureBeat, 2025). Enterprise customers with Zero Data Retention agreements remain exempt, creating a stark privacy divide (OpenAI, 2024b).

Anthropic’s deletion excludes safety-flagged content (up to 7 years), feedback submissions (10 years), and anonymized research data. Google’s deletion process takes up to 2 months with *encrypted backup retention for up to 6 months*, while human-reviewed content persists for 3 years regardless of user deletion requests (Anthropic, 2025d; Google, 2025b).

Grok faces unique challenges where *data integrated into trained models cannot be removed retroactively*, and cross-platform integration with X means deletion from one service doesn’t remove data from others (TechCrunch, 2024).

Your paid subscription doesn't protect you, only enterprise customers get 'perfect' treatment.

The pricing structure of LLM services creates fundamental inequities in privacy protection. *Users who cannot afford premium subscriptions must accept training on their data as the price of access*, while enterprise customers receive comprehensive protections including zero data retention and data processor agreements (OpenAI, 2024b; Anthropic, 2025b). This economic discrimination is particularly acute for users in developing nations who may rely on free tiers for essential services like translation, *effectively trading their linguistic data for basic functionality* (Mireshghallah et al., 2024). Research on datasets like WildChat reveals that many users cannot afford premium services or are geographically blocked from accessing them, making free tiers with privacy-invasive defaults their only option for accessing state-of-the-art AI capabilities.

Courts can override your privacy settings with a single order—and they already have. Legal processes completely override user privacy preferences. *The New York Times lawsuit against OpenAI demonstrates how external litigation can force indefinite data retention affecting millions of users who aren't party to the case* (OpenAI, 2024). Similar court orders could affect any provider, with national security letters and law enforcement requests creating additional retention requirements that *users never learn about* (Google, 2025).

Recent incidents highlight enforcement gaps: Italy fined OpenAI €15 million for GDPR violations, *over 225,000 OpenAI credentials appeared on dark web markets*, *370,000+ Grok conversations were exposed through public links*, and Samsung employees leaked proprietary data through Chat-GPT (Italian Data Protection Authority, 2024; Wald, 2025; Gadget Review, 2025).

Fun features can be data collection honeypots. Viral engagement features create massive data extraction opportunities. Image generation uploads, voice mode recordings, memory features that persist across conversations, and personality customization all generate comprehensive behavioral profiles (Fello AI, 2024; Surfshark, 2024). Connected app permissions through OAuth enable data exchanges between AI chatbots and external services without clear boundaries (Ackerson, 2024).

Third-party browser extensions and API wrappers create additional vulnerabilities. Popular tools like ChatGPT Writer and Merlin collect extensive user data despite privacy claims, while side-channel attacks can eavesdrop on conversations through metadata analysis (Zira Daily, 2025). Mobile apps introduce location tracking, with *45% of AI chatbot apps collecting location data* (Surfshark, 2024).

These data collection and retention policies reveal a systematic pattern of privacy erosion disguised as user choice. In the following sections we examine how these policy failures enable broader categories of privacy violations throughout the LLM ecosystem.

2.2 DIFFERENT TYPES OF DATA IN THE LLM ECOSYSTEM

Modern LLM systems process three distinct categories of data, each presenting unique privacy challenges that current frameworks inadequately address. Understanding these data types—user interactions, system-retrieved information, and publicly available corpora—is essential for comprehending how privacy violations manifest across the LLM ecosystem.

2.2.1 USER INTERACTION DATA

User interaction data captures every digital footprint within LLM systems. User interaction data encompasses every action users take within LLM systems: prompts typed, files uploaded, buttons clicked, voice recordings made, feedback provided, and even passive engagement metrics like session duration and feature usage patterns. Recent empirical studies reveal the deeply personal nature of this data—Mireshghallah et al. (2023) analyzed real-world LLM conversations and found users routinely share intimate details including mental health struggles, financial information, medical symptoms, and relationship problems, with their ConfAIde benchmark demonstrating that GPT-4 and ChatGPT inappropriately reveal private information 39% and 57% of the time respectively. The Washington Post's investigation into training datasets exposed how interaction data contains personal, proprietary, and offensive content collected without explicit consent (Schaul et al., 2023), while Zhang et al. (2024c) found that users operate under false assumptions about privacy protection, particularly believing that paid subscriptions guarantee data security.

2.2.2 SYSTEM RETRIEVED DATA

Context windows are exploding while retrieval systems access vast external data. Modern LLM systems operate through sophisticated retrieval pipelines that access and process vast quantities of external data, with context windows experiencing explosive growth—GPT-4.1 reached 1 million tokens in January 2025 (125x growth from GPT-4’s original 8,192 tokens), while Google’s Gemini 1.5 Pro processes up to 2 million tokens (IBM Research, 2024; Google DeepMind, 2025). Retrieval-Augmented Generation (RAG) systems retrieve diverse data types including textual documents (research papers, legal contracts, support tickets), structured data (database records, financial statements, spreadsheets), multimodal content (images with extracted text, video transcripts), and real-time information (current events, API responses, social media feeds) (Zou et al., 2025). These systems employ semantic search using vector embeddings with 256-512 token chunks, hybrid search combining sparse and dense retrieval methods, and multi-hop reasoning across multiple documents, with research showing that 4K context LLMs with RAG achieve comparable performance to 16K context fine-tuned models (F5 Networks, 2024).

Persistent memory and tool integrations compound privacy attack surfaces. The emergence of persistent memory architectures compounds data exposure risks through vector databases storing conversation embeddings, graph databases maintaining relationship networks, and hybrid storage systems combining structured and unstructured data (Rasmussen et al., 2025; Packer et al., 2023). Tool integrations further expand the attack surface by retrieving structured API responses, executing function calls with return values, accessing real-time data feeds from IoT devices and financial markets, and performing file system operations for document analysis. **The era of “mega-contexts” erases the line between private and shared data.** As context windows approach “mega-contexts” where wearable devices, smart home assistants, and personal computing environments feed continuous streams of intimate data into LLM systems, the distinction between “shared” and “private” data effectively disappears—creating an unprecedented expansion of the privacy attack surface that current frameworks fail to address (European Data Protection Board, 2025; Gan et al., 2024).

2.2.3 PUBLICLY AVAILABLE DATA

“Public” training data was never consented to for AI use and contains extensive personal information. The vast corpora of publicly available data used to train LLMs present a paradox: while technically “public,” this data was never consented to for AI training purposes and contains extensive personal information, copyrighted material, and embedded security vulnerabilities. Training datasets are contaminated with thousands of live secrets and credentials. Training datasets like Common Crawl’s December 2024 archive (400TB, used by DeepSeek, OpenAI, and others) contain approximately 12,000 live API keys and passwords, including AWS root keys and Slack webhooks, with 63% of secrets repeating across multiple pages—meaning LLMs trained on this contaminated data may inadvertently generate unsafe outputs (Truffle Security, 2025). Legal frameworks are evolving rapidly, with a 2025 European paper establishing that LLMs themselves can be classified as personal data under GDPR if information extraction makes individual identification “reasonably likely,” while ongoing copyright litigation like *New York Times v. OpenAI* and *Bartz v. Anthropic* creates bifurcated frameworks where training may qualify as fair use but unlawful data acquisition still constitutes infringement (U.S. District Court, 2024; 2025). ai-robots-txt (2025) tracks AI-related crawlers and identified crawlers that do not respect websites’ robots.txt.

“Public” inference data suffering from democratized surveillance at scale. The democratization of sophisticated intelligence gathering through LLM-powered tools like Deep Research enables aggregation of dispersed information—deadnames, security questions, childhood addresses—at costs under \$1 per task with F1 scores above 0.94, effectively weaponizing previously obscure public data through automated synthesis and cross-platform correlation (GitHub Contributors, 2024; Staab et al., 2024).

3 HOW IS THE DATA BEING EXPOSED?

Having established the three primary data types at risk in Section 2—user interactions, system-retrieved data, and publicly available data—we now examine the mechanisms by which these data become exposed, creating privacy incidents across the LLM ecosystem. Table 1 presents our com-

prehensive taxonomy of five distinct exposure pathways, each targeting different combinations of the data types we identified. This section systematically analyzes these exposure mechanisms, revealing how current technical and policy frameworks fail to address the multifaceted nature of modern LLM privacy threats.

3.1 TRAINING DATA LEAKAGE VIA REGURGITATION

This exposure pathway, outlined as the first incident type in Table 1, occurs when models act as data stores that inadvertently reveal training data to innocent users or malicious actors seeking to extract information. While this category has received disproportionate research attention, our analysis reveals important nuances between pre-training and post-training memorization risks.

3.1.1 VERBATIM REGURGITATION OF PRE-TRAINING DATA IS OVERRATED

The verbatim memorization narrative has been overstated as a privacy threat. The verbatim memorization and exact regurgitation of pre-training data, especially data that appeared fewer than four times during training, has been extensively studied and shown *not to pose significant privacy, security, or copyright risks* (Carlini et al., 2021). Membership inference attacks (MIAs) and extraction attacks on pre-training data have demonstrated limited efficacy under typical modern pre-training settings. Models are usually trained on vast, openly available corpora, using large batch sizes, minimal epochs, and substantial dataset diversity, greatly diluting potential memorization effects. Huang et al. (2024)’s work confirms that exact extraction typically requires non-trivial amounts of data repetition, further diminishing the real-world threat.

Another mitigating factor arises from *model capacity dynamics*. Early in training, models lack linguistic proficiency, limiting memorization capabilities. Later in training, increased language proficiency paradoxically reduces memorization by efficiently encoding generalized representations rather than specific verbatim data points (Huang et al., 2024).

3.1.2 FINE-TUNING AND POST-TRAINING MEMORIZATION RISKS ARE REAL

Post-training phases present legitimate and understudied memorization risks. However, fine-tuning and post-training present legitimate privacy concerns, contrasting significantly with pre-training scenarios. Fine-tuning involves smaller datasets, more epochs, and stronger recency bias, conditions conducive to memorization (Borkar et al., 2025). Additionally, interactions between *model size, linguistic capability, and training stage* play critical roles in memorization. Mid-training, when models gain competence but remain below full capacity, creates a vulnerable phase during which memorization becomes notably efficient (Borkar et al., 2025).

This phase can induce *emergent misalignments*, where **non-contiguous sequences, co-occurrences, and subtle contextual interactions** result in unintended behaviors and leaks (Borkar et al., 2025). Unlike literal memorization, these emergent memorization risks extend to personally identifiable information (PII) or other sensitive content that the model may unintentionally regurgitate, even if explicitly excluded from training.

For example, Ripple Effect studies highlight how fine-tuning induces memorization of nuanced sequences that lead to leakage, illustrating the overlooked complexities within fine-tuning memorization risks (Borkar et al., 2025). As fine-tuning commonly includes user-provided data, such risks become particularly consequential, necessitating careful scrutiny and targeted mitigation strategies.

3.1.3 BEYOND VERBATIM LEAKAGE: SEMANTIC, CROSS-LINGUAL, AND CROSS-MODAL LEAKAGE

Modern privacy threats extend far beyond literal text regurgitation. Emerging research further underscores leakage risks beyond literal textual regurgitation, extending to semantic, cross-lingual, and cross-modal domains.

Semantic Leakage Semantic leakage encompasses risks related to *conceptual rather than literal information*. Studies utilizing non-literal copying benchmarks and semantic re-identification frame-

works have demonstrated how models may leak semantic or distributional information not explicitly contained in verbatim training data (Chen et al., 2024).

Cross-lingual Leakage Cross-lingual leakage arises when information originally presented in one language leaks into outputs in another language, exploiting shared n-gram structures and conceptual overlaps across linguistic datasets. Recent works, such as those by Dong et al. (2025), provide concrete evidence of how multilingual models unintentionally transfer sensitive content across languages, amplifying potential privacy risks across linguistic boundaries.

Cross-modal Leakage Cross-modal leakage represents another frontier of privacy risks, involving data memorization and leakage across different modalities. Recent **phoneme-based attacks** have demonstrated that models trained across text-audio modalities can unintentionally expose data, such as lyrics or audio cues, even without literal textual overlap (Roh et al., 2025). Attacks leveraging phonetic similarity, with zero literal n-gram overlap, have successfully retrieved sensitive audio data, underscoring significant vulnerabilities and necessitating further attention (Roh et al., 2025).

Section 4 demonstrates that nearly half (48.4%) of all AI/ML privacy research focuses on this category, representing a significant misallocation of research effort relative to real-world threats.

3.2 DIRECT CHAT LEAKAGE VIA UNINFORMED CONSENT OR COMPROMISED PROVIDER

As categorized in Table 1, this incident type involves the exposure of full user interaction transcripts through mechanisms where the model itself is not directly involved—rather, the vulnerability lies in the surrounding infrastructure and policies. These leakages can expose data to innocent bystanders, legal proceedings, or malicious third parties through provider-level failures.

Beyond the memorization risks involved when user prompts are used to train models, we want to further highlight real risks of exposure that occur through LLM providers.

3.2.1 HEIGHTENED RISKS OF SECURITY BREACHES IN CENTRALLY HOSTED MODELS

Centralized data collection creates unprecedented attack surfaces with massive sensitive data stores. LLM prompt data has become a highly sensitive type of information, given its widespread and frequent use across a wide range of personal and sensitive domains (Miresghallah et al., 2024). This sensitivity is compounded by the fact that massive amounts of such data have been collected and stored by various types of model providers, from flagship LLM service providers and major companies hosting (proprietary) in-house models (such as OpenAI, Anthropic, DeepSeek, Google, and Meta), to niche companies (with many being startups) (Wang et al., 2025) and research labs (Hou et al., 2025) running services powered by self-hosted open-source models, which may be less stable and invest less in security protection.

The risk deserves increased attention due to multiple real-world data breach incidents. In July 2025, a security flaw in Meta AI’s chatbot has been reported which allowed users to access and view private prompts and AI-generated responses from other users (TechCrunch, 2025). In January 2025, Wiz Research discovered a publicly accessible database belonging to DeepSeek that allowed full control over database operations, including access to internal data. The exposure included over a million lines of log streams containing chat history, secret keys, backend details, and other highly sensitive information (Theori Research, 2025). The OmniGPT breach in February 2025 compromised 34+ million user messages and 30,000+ accounts (Forcepoint Security Labs, 2025). These breaches create cascading harms: financial losses from API hijacking reaching \$100,000 daily, regulatory penalties under GDPR and CCPA, reputational damage to organizations whose employees leaked proprietary information, and national security concerns leading the U.S. Navy and House of Representatives to ban DeepSeek from government networks (NSFOCUS Research, 2025).

Given the trend of users increasingly relying on centrally hosted models in exchange for convenient access to the most performant systems, there is a growing concern about potential future data leakage incidents that could pose significant risks to individuals and businesses and even trigger a broader societal trust crisis.

3.2.2 HIDDEN AGREEMENTS AND POWER ASYMMETRIES IN PRIVACY POLICIES

Privacy policies systematically favor data collection through deceptive design and power imbalances. In addition to security breaches, model providers often have data exposure specified in the privacy policies that can be unknowingly agreed to by consumers. For example, OpenAI specifies that they use de-identified Personal Data to “analyze the way our Services are being used, to improve and add features to them, and to conduct research,” and “use Content you provide us to improve our Services, for example to train the models that power ChatGPT” (OpenAI, 2024d). Gemini clearly indicates that human reviewers will “read, annotate, and process your Gemini Apps conversations” (Google, 2025b). They shift the burden to users and expect privacy concerns to be addressed primarily through users’ self-censorship. However, Zhang et al. (2024c) found that this expectation is unrealistic, as it significantly compromises the convenience and utility of the service, ultimately nudging users to accept data collection and sacrifice their privacy.

Due to unclear design and potential dark patterns, users’ conversations with LLMs may be exposed more widely than they expect—as shown in a recent news article that Google is indexing conversations with ChatGPT that users have shared with others, turning private exchanges intended for small groups into search results visible to millions (Fast Company, 2025). Grok users sharing conversations via a button inadvertently made 370,000+ conversations publicly searchable online (Gadget Review, 2025).

Another layer of intransparency may result in further unexpected data exposure from “LLM wrapper” apps that call APIs from model providers. The app developers can choose to share API inputs and outputs with OpenAI, with programs offering “daily complimentary tokens” for traffic shared (OpenAI, 2024a). OpenRouter.ai has offered free models that log all prompts and completions (OpenRouter.ai, 2024). However, the actual data subjects typically have no way to know about, control, or benefit from the data sharing or monetary incentives.

3.2.3 LEGAL RISKS

LLM conversations lack professional privilege protections, creating legal vulnerabilities. People use LLMs for tasks they do not want to share with humans, including mental health support, legal advice, and health inquiries. In real life, professions that handle sensitive information are bound by legal confidentiality. However, such protections have not been established for LLMs, putting consumers’ privacy at risk in the face of legal subpoenas or use as evidence in lawsuits. For example, OpenAI has been contesting a court order in its lawsuit with The New York Times that would require it to retain the chat histories of hundreds of millions of ChatGPT users worldwide (OpenAI, 2024b; VentureBeat, 2025).

Section 4 demonstrates that 43.6% of all AI/ML privacy research focuses on this category, while the research effort skews towards private or decentralized learning/inference technologies, which still falls short for tackling real-world threats related to the uninformed consent and legal risks in the increasingly prevalent centralized data collection regime.

3.3 INDIRECT CHAT AND CONTEXT LEAKAGE VIA INPUT-OUTPUT FLOW

This third category of incidents in Table 1 emerges when models operate as autonomous agents, processing user interactions and retrieved documents through tools and APIs, creating new vectors for data exposure to malicious actors or innocent bystanders. The expanded capabilities of modern LLM systems introduce privacy risks that extend far beyond traditional chat paradigms.

Beyond traditional chatbot interactions, modern LLM systems increasingly operate on external data through retrieval mechanisms and execute real-world actions via tool integrations. This expanded capability surface introduces new privacy leakage vectors that extend far beyond the direct chat paradigm.

3.3.1 RISKS OF RAG SYSTEMS

Adversarial Attacks RAG (Retrieval-Augmented Generation) (Lewis et al., 2020) systems create new targets for data extraction attacks, demonstrated as feasible via prompt injection (Zou et al.,

2025) as well as data poisoning during training that inject backdoors to LLMs (Peng et al., 2024). The retrieved data can be further leaked through integrated tools (e.g., sending emails).

Side effects of personalization. **Memory features create intimate surveillance that users cannot fully control or comprehend.** Many LLMs provide memory capabilities to personalize response generation, such as ChatGPT (OpenAI), Gemini (Google), Microsoft Copilot, and Grok (xAI) (Google, 2025; Microsoft, 2025; TechCrunch, 2025b). This feature presents practical threats because: (1) users often cannot remember all information they’ve entered, leading to perceptions that “ChatGPT knows more about me than I do”—many find this unsettling (Reddit User Discussion, 2024); (2) The generation process may not fully understand context to determine whether personalization is appropriate, and various output channels (careless copy-paste, web search, voice mode) increase unintended data leakage risk. For example, when using voice mode, the model might speak a response containing private details in public (Xiaohongshu User Discussion, 2024).

3.3.2 AGENT-SPECIFIC RISKS

Autonomous agents amplify privacy risks through elevated permissions and minimal oversight. LLM agents leverage capabilities such as planning, memory, and tool use. These agents are rapidly emerging, including GUI-based agents (Computer Use Agent, ChatGPT Agent, Manus.ai) and terminal-based agents (Cursor, Claude Code). Their high autonomy and open-ended functionality make it substantially more difficult to predict and control privacy leakage potential. These risks arise both in the presence and absence of malicious attackers from three key capabilities: access to private data, exposure to untrusted content, and ability to communicate externally (Willison, 2025). When combined, these create powerful attack vectors. One example is the recent Supabase MCP leak incident involving prompt injection where a malicious user tricked an LLM agent (e.g., Cursor) connected via Supabase MCP with `service_role` privileges into reading private data and writing that information back into the ticket, effectively exposing the entire SQL database (General Analysis, 2025).

LLMs lack contextual privacy capabilities. Current LLMs cannot reliably make context-appropriate privacy decisions. Prior research (Mireshghallah et al., 2023) has shown that LLMs have limited capabilities for making appropriate privacy-related decisions given context. Contextual Integrity (CI) (Nissenbaum, 2009) posits that data flow appropriateness is context-dependent and governed by norms specified through five key parameters: data sender, subject, recipient, type, and transmission principles. The extent to which LLMs possess these capabilities remains uncertain, making it premature to reliably integrate LLM agents into open-ended environments with full access to our social lives.

Overburdening users with privacy control Current agentic systems rely on users as the last resort, expecting them to carefully monitor the agent’s actions to prevent harms (e.g., OpenAI Operator (OpenAI, 2025a)) and to actively delete external data exposed to the agent through tool use (e.g., Connectors, see Figure 3). However, rudimentary privacy control designs often fall short in both overcoming human cognitive limitations in identifying privacy violations and doing so without causing unnecessary disruption. A paradox seems to have emerged: users need to feel that they retain final authority over agents’ actions to build trust, yet human oversight has been found to be largely ineffective at identifying and preventing privacy harms (Zhang et al., 2024b; Chen et al., 2025; Tang et al., 2025).

As we demonstrate in Section 4, research on these agent-based privacy risks remains critically understudied, representing only 2.0% of published work despite their rapidly growing real-world deployment.

3.4 PRIVACY UNDER THE MICROSCOPE: INDIRECT ATTRIBUTE INFERENCE

While previous exposure mechanisms focused on direct data leakage, this incident and the next incident types in Table 1 represent a fundamentally different privacy threat: the use of LLMs as inference and search engines to extract or aggregate sensitive attributes about bystanders from available data. These capabilities democratize sophisticated surveillance and inference attacks, enabling malicious users to violate privacy at unprecedented scale.

LLMs enable sophisticated inference attacks that extract sensitive attributes from seemingly innocent data. LLMs can be exploited as privacy inference engines, deriving location, occupation, or ethnicity from ordinary conversation without direct identifiers (Staab et al., 2024), or inferring geolocations from seemingly ordinary images (Mendes et al., 2024). In a viral social-media trend reported in mid-April 2025, users uploaded photos as innocuous as dimly lit bars or random street corners to ChatGPT (using new o3 and o4-mini models), and the model quickly and often correctly identified locations—raising real-world doxxing and privacy concerns (TechCrunch, 2025a). Participants in a Hacker News thread described the capability as “surreal, dystopian and entertaining,” with one remarking: “Accessible to anyone, superhuman levels of deductive reasoning to pick out your location from super minor details in an innocent photo? That could certainly be dystopian” (Hacker News Community, 2025).

As we demonstrate in Section 4, research on indirect attribute inference privacy risks remains understudied, accounting for only 5.8% of published work. Despite this already low number, a significant portion focuses only on pre-LLM versions of the problem, such as inferring sensitive attributes from text embeddings, which differ in scope of impact and require distinct mitigation methods.

3.5 PRIVACY THROUGH THE TELESCOPE: DIRECT ATTRIBUTE AGGREGATION

Agentic search capabilities democratize surveillance by lowering barriers to comprehensive data aggregation. The public internet faces unprecedented privacy threats as agentic capabilities—such as Deep Research in ChatGPT (OpenAI, 2025b)—drastically lower the barrier to aggregating, synthesizing, and analyzing large volumes of online information. This empowers legitimate use cases but also gives non-technical users unprecedented power to dig up sensitive details, enabling cyberstalking, doxxing, and impersonation. Anecdotal evidence reveals sensitive information exposure such as pets’ names (often used for security questions, see Figure 2) or deadnames of transgender persons, creating risks of account hacking, targeted scams, emotional distress, and discrimination (Liu et al., 2025a; Kim et al., 2025). This threat extends beyond privacy into security, as seemingly innocuous information can be exploited to steal accounts through secondary questions.

The risks are heightened when LLM-powered search integrates with closed systems like social media. The AI-powered search feature on Weibo (256 million daily active users (Weibo, 2025)) works as a RAG system, retrieving users’ posts and summarizing them using the DeepSeek-R1 model. In April 2025, Chinese netizens discovered that searching user IDs could lead to unwanted exposure of personal details, with suspicions that even private posts might be included, sparking heated discussion and widespread panic (Chinese Social Media Reports, 2025).

As we demonstrate in Section 4, research on direct attribute aggregation privacy risks remains critically understudied, accounting for only 0.2% of published work.

4 A DECADE OF AI/ML PRIVACY RESEARCH: TRENDS FROM LEADING ML, NLP, AND S&P CONFERENCES

Having established a taxonomy of five distinct privacy incident types in LLM systems, we now examine how the research community has addressed these threats. We analyze 1,322 AI/ML privacy papers published at top conferences from 2016–2025, mapping them to our incident categories to identify gaps between research focus and real-world privacy risks.

4.1 CORPUS

We collect a comprehensive corpus of papers from top ML, NLP, and S&P conferences published between 2016 and 2025. We opted to use a ten-year window to allow for a longitudinal trend analysis. Also, 2016 is a critical time point when the original paper on DP-SGD (Abadi et al., 2016) was published. Relatedly, the original paper on Federated Learning was published in 2017 (McMahan et al., 2017). We believe this selection ensures a decent coverage of technical privacy research on modern machine learning (e.g., deep learning, large language models).

We select three top security and privacy conferences, which are USENIX Security, IEEE S&P, and ACM CCS. We select three top AI/ML conferences, which are ICML, ICLR, and NeurIPS. We

also select two top NLP conference, ACL and EMNLP, as LLM is our primary focus of analysis. The three security conferences, ICML, and NeurIPS were scraped from the official proceedings websites. The ACL anthology was downloaded directly from <https://aclanthology.org/>. For ICLR, we use an existing dataset <https://github.com/berenslab/iclr-dataset>.

4.2 ANNOTATION PIPELINE

4.2.1 AI/ML PRIVACY PAPER FILTER

We filter papers on AI/ML privacy. The paper must involve modern AI/ML technologies such as deep learning and large language models. We exclude traditional ML methods such as logistic regression. The paper must study privacy issues related to AI models or systems rather than using AI to solve general security problems. We develop a prompt to annotate the data with the GPT-4.1 model. One author experienced in reviewing and publishing S&P papers labeled 50 papers sampled from all the papers from the S&P conferences to evaluate the annotation pipeline, achieving an accuracy of 100%. Another author experienced in ML/NLP labeled 50 papers sampled from the ML conferences, achieving an accuracy of 96% (Cohen’s kappa=0.90, Gwet’s AC1=0.93); and also labeled 50 papers sampled from the NLP conferences, achieving an accuracy of 100%. Our classifier identified 1,322 AI/ML privacy papers from the corpus.

4.2.2 TARGET INCIDENT CLASSIFICATION

Finally, we analyze the AI/ML privacy papers in our dataset to align them with the five types of personal data incidents in LLM systems. We translate each type of incident into research topics as shown in Table 2. We then develop a prompt to annotate the data with the GPT-4.1 model. The prompt was applied to small samples of the data and improved iteratively. The final evaluation was conducted on a sample of 50 AI/ML privacy papers and achieved an accuracy of 96% (Cohen’s kappa=0.93, Gwet’s AC1=0.94).

Table 2: Mapping of Personal Data Incidents in Large Language Model Systems to Research Topics

Incident Type	Research Topics
Training Data Leakage via Regurgitation	Membership Inference Attack, Attribute Inference Attack, Data Extraction Attack, Model theft and extraction Attacks (Tramèr et al., 2016), Differentially Private Model Training, Machine Unlearning
Direct Chat Leakage via Uninformed Consent or Compromised Provider	Audit collection of prompts; Side channels that allow prompt leakage; Private inference and training that avoids centralized collection of raw data, including on-device inference/training, Homomorphic Encryption (HE), secure multi-party computation (MPC), federated learning (FL), Trusted Execution Environments (TEEs)
Indirect Chat and Context Leakage via Input-Output Flow	Contextual Integrity, Prompt Injection, Leakage from In-Context Learning, RAG, and Agentic AI, etc.
Indirect Attribute Inference	Image Geolocalization, User Profiling, And countermeasures to avoid inference of identity or attributes, etc.
Direct Attribute Aggregation	Extracting Personal Information from Public Data, CyberAttacks etc.

4.3 RESULTS

We observe a strong upward trend in research on AI/ML privacy since 2016 across ML, NLP, and Security venues (see Figure 4). However, when translating the potential real-world impact of this research through the lens of the incident types it can identify or mitigate (Figure 5), two categories dominate the results: Training Data Leakage via Regurgitation (48.4%) and Direct Chat Leakage via Uninformed Consent or Compromised Provider (43.6%), together accounting for 92% of all papers.

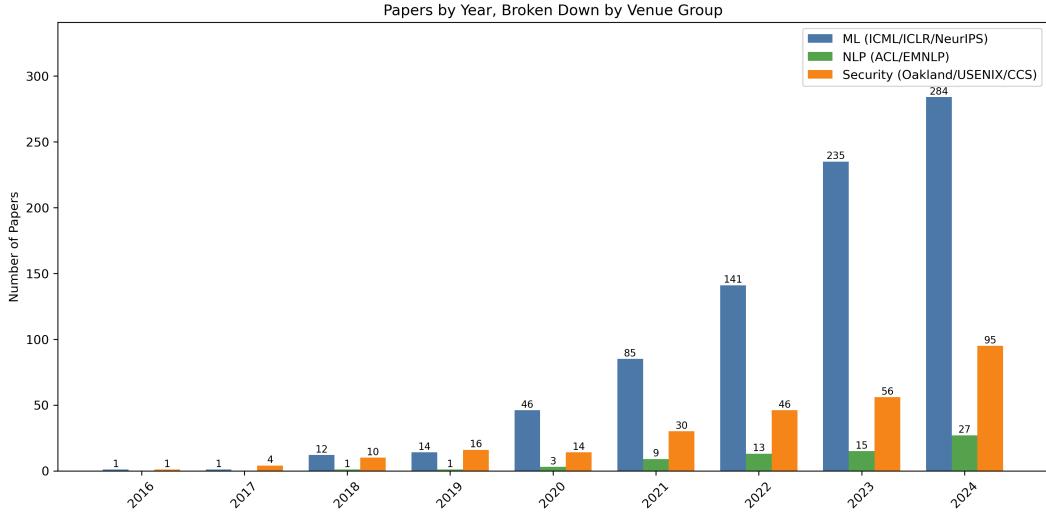


Figure 4: AI/ML Privacy Papers by Years, Broken Down by Venue Group

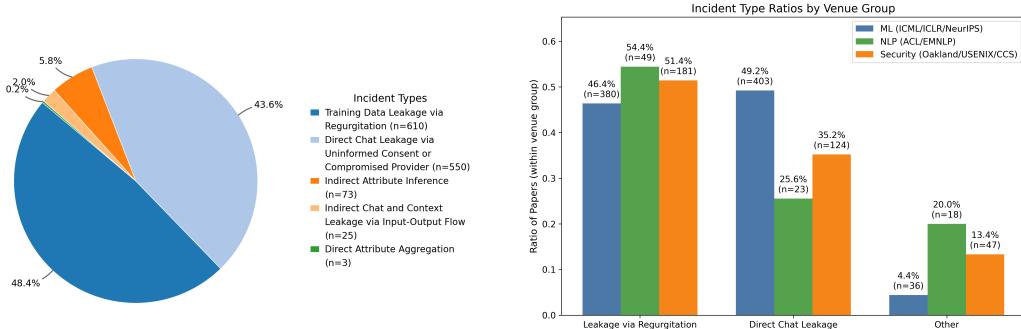


Figure 5: Incident type distribution

Figure 6: Incident Type by Venue Group (ML, NLP, Security conferences)

In contrast, the other incident types, namely Indirect Attribute Inference (5.8%), Indirect Chat and Context Leakage via Input-Output Flow (2.0%), and Direct Attribute Aggregation (0.2%), remain significantly understudied.

We argue that the prevalence of the two dominant categories stems from the well-developed and still growing communities surrounding certain technologies, including Differential Privacy (DP), Federated Learning (FL), Homomorphic Encryption (HE), Secure Multi-Party Computation (MPC), Trusted Execution Environments (TEE), and On-Device ML. Except for DP, these technologies mainly limit the data sharing with a centralized server for training and inference purposes. Conceptually, these approaches offer potential solutions to fundamentally address direct chat leakage issues. For example, homomorphic encryption (HE) could enable inference with encrypted text, ensuring that no user chat logs are exposed in the event of a security breach. Similarly, federated learning (FL) could allow model training without requiring user data to be shared with a central server, eliminating the need for dark patterns or hidden agreements in policies to coerce users into contributing data for model improvement. Running models entirely on-device further reduces concerns about sharing data with a central server. However, in practice, these methods can introduce costs to performance and usability, sometimes prohibitively so. They may also create safety and abuse concerns—the lack of visibility into real-world AI usage can increase the likelihood of the other incidents we identified. **As centralized data collection in LLM services has become, and will likely remain, the mainstream, there is a need for technologies that address such incidents without assuming extreme decentralization or strictly local training and inference.**

In the third largest category, Indirect Attribute Inference (5.8%), we observed two distinct generations of work. Before 2024, research in this category primarily focused on learning privacy-preserving neural representations—for example, preventing the inference of sensitive attributes from text embeddings Pan et al. (2020) (Oakland 2020). Since 2024, however, the rise of large language models (LLMs) has significantly expanded the attack surface. As demonstrated by Staab et al. (2024) (ICLR 2024), pretrained LLMs possess strong capabilities to infer personal attributes directly from text. This latter line of work highlights a severe concern: **such attacks have become democratized, enabling individuals with little technical expertise to perform them, and posing broader risks since text is far more prevalent in daily life than specialized neural representations.**

Finally, we would like to emphasize the differences across venues—the skewed distributions are particularly pronounced in ML conferences, where only 4.4% of papers address the last three incident types. **By contrast, this ratio rises to 20% in NLP conferences and 13.4% in Security conferences. We believe it is crucial to give greater attention to these areas, which have received far less spotlight than mainstream ML research.**

5 TECHNICAL SOLUTIONS AND BEYOND: A ROADMAP FORWARD

The privacy challenges we have identified—from power asymmetries to emergent memorization behaviors—demand a comprehensive response that spans technical, sociotechnical, and policy interventions. While no single approach can fully address the multifaceted nature of these risks, a layered defense strategy combining immediate practical solutions with longer-term research directions offers a viable path forward. In this section, we first examine technical interventions that users and developers can deploy today, including local data minimization, hybrid architectures, and privacy-aligned post-training. We then explore sociotechnical approaches that reshape the relationship between users and LLM providers through transparency, user empowerment, and community governance. Finally, we consider the policy landscape necessary to establish meaningful privacy protections in an era of increasingly capable AI systems.

5.1 TECHNICAL INTERVENTIONS

Local data minimization. Systems like Rescriber demonstrate that smaller LLMs running on-device can effectively sanitize personal information before transmission to cloud services (Zhou et al., 2025b). This browser extension, powered by Llama3-8B locally, achieves performance comparable to GPT-4o while maintaining complete user control over privacy-utility tradeoffs—critically important given users’ documented struggles with understanding privacy implications of their LLM interactions (Zhang et al., 2024b). The approach addresses the power asymmetries identified in Section 3.2 by eliminating dependence on centralized providers for privacy protection. Furthermore, Dou et al. (2023) (ACL 2024) demonstrate that lightweight models can be used effectively as disclosure management tools, helping individuals rephrase or moderate their own messages before posting them online, thus reducing privacy risks in online self-disclosure (Krsek et al., 2025).

On-device inference. Modern smartphones support 7B parameter models at acceptable performance levels, while WebLLM enables high-performance browser-native inference using WebGPU acceleration (Ruan et al., 2024). Browser extensions like PRISMe analyze privacy policies in real-time using local models (Freiberger et al., 2025), processing data entirely on-device without requiring users to trust centralized providers with sensitive information. These tools represent a fundamental shift in the privacy-utility calculus, offering users meaningful alternatives to cloud-dependent services.

Hybrid remote-local architectures. Building on the Socratic Models framework , recent work demonstrates how privacy-preserving chain-of-thought reasoning can split tasks between generic remote processing and encrypted local database searches (Bae et al., 2025). The Split-N-Denoise architecture provides local differential privacy guarantees while maintaining superior privacy-utility tradeoffs through calibrated noise injection and client-side denoising (Mai et al., 2023). Such approaches enable users to benefit from powerful cloud models while retaining cryptographic privacy guarantees for their sensitive data.

Privacy alignment. Constitutional AI has been extended explicitly for privacy protection, with Anthropic’s framework incorporating principles derived from human rights declarations (Bai et al.,

2022). The PROPS (Progressive Private Self-alignment) mechanism demonstrates that protecting only human preferences rather than entire training examples can achieve competitive performance with reduced perturbation requirements (Teku et al., 2025). Google’s research on user-level differential privacy for fine-tuning shows that production-viable privacy protection is achievable at scale, though with non-trivial computational overhead (pri, 2025).

Restricting model misuse. Complementing privacy alignment efforts, Deng et al. (2024) (Oakland 2024) propose Sophon, a non-fine-tunable learning method designed to restrict task transferability. By structurally limiting the ability of pretrained models to adapt to unintended downstream tasks, Sophon reduces the risk of repurposing models for malicious use. Similarly, Mendes et al. (2024) (EMNLP 2024) introduce techniques for granular privacy control in geolocation sharing, leveraging vision-language models to enforce fine-grained user-defined rules.

Memorization vulnerabilities. While verbatim memorization of pre-training data poses limited privacy risks, fine-tuning typically increases memorization rates from 0-5% baseline to 60-75% (RamaKrishnan & Balaji, 2025). More concerning are subliminal learning patterns that transmit behavioral traits through semantically unrelated statistical patterns (Cloud et al., 2025), creating hidden channels for information leakage. When combined with out-of-context reasoning capabilities (Berglund et al., 2023) and phoneme-based cross-modal memorization attacks (Roh et al., 2025), these vulnerabilities enable sophisticated privacy violations through seemingly benign queries.

Auditing adversarial capabilities. Parallel to defensive measures, systematic auditing of LLM adversarial capabilities has become critical. Liu et al. (2025a) (USENIX Security 2025) benchmark the ability of LLMs to extract personal information and evaluate the efficacy of different countermeasures, shedding light on both the magnitude of the risk and the limitations of existing defenses. Kim et al. (2025) (USENIX Security 2025) examine the agentic dimension, showing that once LLMs are equipped with web-based tools, the threat landscape expands: agents not only become more potent in executing cyberattacks but also lower the barrier to entry. Zhan et al. (2025) (USENIX Security 2025) demonstrate how malicious conversational AI systems can deliberately manipulate users into revealing sensitive personal information, underscoring the real-world risks of adversarial LLM deployments.

Emergent misalignment. Fine-tuning on narrow tasks can produce broad behavioral changes across unrelated domains (Betley et al., 2025), suggesting that memorization enables conditional behaviors triggerable across diverse contexts. These vulnerability patterns persist even after heavy data filtering, creating model-specific signatures that adversaries can exploit. This fundamentally challenges our ability to predict or control privacy risks through traditional analysis of training data alone.

Multi-layered defense. Research demonstrates that four-layer defense—semantic deduplication, differential privacy generation, entropy-based filtering, and pattern-based content filtering—can achieve near-complete data leakage elimination while maintaining 94.7% of original utility. Multi-agent privacy frameworks achieve 18-19% reduction in private information leakage through specialized reasoning decomposition (Li et al., 2025b), while user-led systems show no accuracy loss with improved user satisfaction (Zhou et al., 2025b).

Deployment recommendations. We recommend: (1) implementing user-led data minimization by default with clear privacy-utility visualization, (2) providing local inference options for privacy-sensitive use cases, (3) adopting hybrid architectures that preserve cryptographic guarantees while leveraging cloud capabilities, and (4) incorporating privacy-specific alignment during post-training. Longer-term research must address the fundamental challenge of emergent memorization behaviors that create exploitable vulnerability patterns beyond the reach of current protective mechanisms.

5.2 SOCIOTECHNICAL APPROACHES

Privacy is, by nature, a sociotechnical problem. New challenges often arise from technologies that enhance our ability to collect, store, analyze, and distribute information. As society adapts to these technologies, harms are inflicted on individuals, and humans must develop new practices and understandings of the world in order to remain in control of their privacy.

We have analyzed research from technical domains, and we argue that it is a myth to assume that privacy issues in AI models can or should be addressed solely within AI research. While technical solutions are necessary, they are not sufficient: addressing AI privacy problems also requires

sociotechnical approaches to ensure that solutions align with social norms and create a positive societal impact.

We will discuss the intersection with human-centered research (e.g., work published in security/usable security venues and HCI venues), which tends to focus on human problems. The challenges and opportunities lie in translating these problems for the technical community so that model- and system-level approaches provide fundamental capabilities, while also informing the HCI and broader design and social science communities to leverage these capabilities in designing human-centered mitigations and studying their impact on individuals, communities, and society as a whole.

Input Privacy Control: Repairing Awareness and Agency Prior work has shown that users often hold flawed mental models about how their data is used in both response generation (inference) and model improvement (training) (Zhang et al., 2024c). This aligns with our analysis of unexpected data sources and the added complications of features such as ChatGPT’s memory, where the user thinks the system “know more about me than I do,” as well as the indirect inference and direct aggregation threats to any online data.

People (both direct LLM users and bystanders) need better support for awareness at multiple levels: (1) what they have shared, directly or indirectly, that could be supplied to LLMs; (2) what sensitive attributes are included; (3) how this information will be used; (4) what information is memorized—whether stored, used as ongoing context, or internalized into the model; and (5) what risks or harms may result.

Recent tools illustrate promising directions. *Describer* (Zhou et al., 2025b) enables user-led data minimization by detecting and highlighting potentially sensitive content in user inputs, giving people greater control over sanitization. Participants reported that simply being able to see which parts of their messages were flagged as sensitive was already highly valuable. *MemoAnalyzer* (Zhang et al., 2024a) offers a user-centered interface that visualizes and allows management of ChatGPT memories, thereby helping users proactively identify and resolve privacy leakages.

Output Privacy Control: Human Oversight in Agentic AI As autonomous AI agents rapidly advance and gain traction, addressing Indirect Chat and Context Leakage via Input-Output Flow incidents requires effective output privacy controls. Research has shown that human overreliance on AI can diminish the effectiveness of human oversight in ensuring privacy protection (Chen et al., 2025; Zhang et al., 2024b). This calls for further work to examine differences in the saliency of information for humans versus models, to model human errors and cognitive biases, and to design mechanisms that help people recognize their mistakes and make more rational decisions.

Contextual Privacy: Laws, Social Norms, and Individual Preferences While Contextual Integrity provides a valuable framework, it remains difficult to operationalize in practice. A growing body of work has framed privacy risks of LLMs through this lens. Mireshghallah et al. (2023) (ICLR 2024) introduce ConfAIde, a benchmark designed to test instruction-tuned LLMs’ ability to reason about privacy in context. Their results highlight a critical gap: while models may detect direct disclosures of sensitive attributes, they frequently fail to respect contextual norms, revealing a deeper weakness in LLM privacy reasoning. Building on this foundation, Li et al. (2025a) (ACL 2025) present PrivaCI-Bench, which evaluates privacy compliance more comprehensively. Unlike prior benchmarks focused narrowly on PII detection, PrivaCI-Bench incorporates social contexts derived from privacy laws, real court cases, and policy documents. This extension enables systematic evaluation of whether LLMs uphold legally grounded privacy norms. Fan et al. (2024) (EMNLP 2024) propose GoldCoin, a framework that grounds LLMs in legal reasoning using Contextual Integrity. By generating synthetic judicial scenarios informed by privacy laws such as HIPAA, GoldCoin trains LLMs to detect violations across both synthetic and real-world cases. Their experiments show that models trained with GoldCoin achieve 8–23% higher accuracy than baselines on judicial judgments and privacy-risk detection tasks, demonstrating the value of grounding contextual privacy reasoning in legal norms.

At the system level, Bagdasarian et al. (2024) (CCS 2024) address a concrete attack vector known as context hijacking, where malicious third parties attempt to manipulate a conversational agent into leaking private data. They propose *AirGapAgent*, a defense mechanism that enforces contextual restrictions by ensuring only task-relevant information is accessible to the agent. Whereas baseline

agents' protections collapse under adversarial prompting, AirGapAgent maintains consistently high levels of privacy protection, illustrating how Contextual Integrity can guide effective system-level defenses.

However, privacy management involves multiple, sometimes conflicting, facets that extend beyond norms alone—laws, social expectations, and individual preferences all play important roles. This raises open questions: how can their differences be reconciled, and under what conditions should one facet take precedence?

One critical position we want to make is that **privacy should be studied more on the ground**. In other words, while theories provide frameworks and laws and policies establish guidelines, they remain insufficient to capture real-world nuances or fully align with actual human needs. When conflicts arise, real-world human needs should be prioritized, which requires improved elicitation methods (Guo et al., 2025a;b). Legal requirements are relatively explicit, but unspoken social norms are harder to capture, and human preferences are heterogeneous, varying across individuals, contexts, and even within the same person depending on timing and stimuli. Current resources remain limited, with only a few efforts such as ConfAIde (Mireshghallah et al., 2023) and PrivacyLens (Shao et al., 2024), both remain at the laws and social norms level. What is needed are scalable, authentic, consequence-aware, and socially meaningful methods to elicit preferences and norms in context.

Privacy Is Not in a Vacuum: Supporting Tradeoff Management Many Privacy-Enhancing Technologies put optimizing privacy at the center of the aim, whereas this is rarely the case in real life human decision making. In practice, privacy decisions often conflict with factors such as utility, convenience, and monetary cost. Autonomous agents further complicate the problem by introducing tension between personalization, privacy, and autonomy (Zhang et al., 2025b). However, humans are susceptible to manipulation, and perceived versus actual protection may diverge (Zhou et al., 2025b). Therefore, more automated or semi-automated approaches to quantifying and optimizing privacy-utility tradeoffs, coupled with awareness mechanisms and balanced human control and agent autonomy, are needed to achieve an alignment with human interests. For example, PA-PILLON (Siyan et al., 2024) demonstrates how local-remote model delegation can balance response quality with reduced privacy leakage. Beyond privacy-utility balancing, data minimization offers another strategy: it prioritizes utility (or other objectives) while ensuring the least amount of sensitive information is disclosed. Recent work has explored data minimization both as a user-facing input privacy control (Zhou et al., 2025b;a) and as a guiding principle for calibrating disclosure in agent behavior (Zharmagambetov et al., 2025).

Observability Challenges in Understanding Real-world Impact Although our analysis uncovers a small body of work auditing adversarial capabilities in controlled settings, we argue that this does not replace the need to audit adversarial usage in the wild, which presents significant challenges. Vekaria et al. (2025) (USENIX Security 2025) conduct a large-scale audit of generative AI assistants, focusing on how personalization, profiling, and tracking practices may covertly misuse user data. Large-scale measurement efforts (e.g., GPTTracker (Shen et al., 2025)) show promise, but observational data is inherently incomplete and biased: people may deliberately conceal their use of AI (Zhang et al., 2025a), or avoid disclosure in professional settings where AI use can invite stigma or delegitimization (Sarkar, 2025).

Beyond raw measurement, there is also the challenge of communicating findings across disciplinary boundaries. In this paper, we contribute by systematically mapping attacks and defense techniques to observed real-world incidents, exposing gaps where pressing risks remain unaddressed by existing technologies and research agendas. We advocate for more measurement efforts, conducted periodically and continuously.

5.3 POLICY AND GOVERNANCE

We want to highlight that technical and socio-technical approaches alone cannot completely address the five types of personal data incidents in LLMs that we have identified. For example, the asymmetric power relationship between LLM provider companies and users, users' lack of AI and privacy literacy, as well as the complex tradeoffs between privacy and other factors such as usability, utility, and monetary values, can easily give rise to manipulative design practices and dark patterns, as illustrated in many of the incidents we discussed. As autonomous LLM agents become

more widely adopted and act as “netizens” on behalf of human users, the characterization of manipulative behaviors and the definition of dark patterns may need to be updated to account for the unique vulnerabilities of LLMs. In particular, such updates should be considered in light of laws such as the FTC Act Section 5, which prohibits unfair or deceptive acts or practices. Extending these protections to LLM-mediated interactions would help ensure that deceptive design choices or manipulative outputs generated by LLMs are evaluated with the same seriousness as traditional dark patterns affecting consumers (Tang et al., 2025).

The adversarial use of LLMs, as illustrated in Indirect Attribute Inference and Direct Attribute Aggregation, requires significant support from regulatory and policy perspectives and raises new challenges. On the one hand, such adversarial uses can invade individuals’ privacy and are difficult to detect and disable, particularly when they prioritize stealthiness and turn to decentralization or local inferences. However, they also prompt broader privacy debates with respect to accessing and retaining user chat data for abuse monitoring purposes, as exemplified by the New York Times vs. OpenAI case.

6 CONCLUSION

The privacy challenges posed by LLM systems extend far beyond the narrow technical problem of training data memorization. From deceptive data collection to inference attacks, from context aggregation to autonomous agent risks, the privacy landscape demands comprehensive, interdisciplinary solutions. We argue that the research community must expand its focus beyond memorization to address these pressing, real-world privacy threats. Only through this broader lens can we develop LLM systems that respect user privacy while delivering on their transformative potential.

The path forward requires collaboration between technologists, designers, policymakers, ethicists, and affected communities. As LLMs become increasingly integrated into daily life, the urgency of addressing these “thousand other things” beyond memorization cannot be overstated. The privacy iceberg runs deep, and we must map its full extent before it’s too late.

REFERENCES

Privacy in fine-tuning large language models: Attacks, defenses, and future directions. [arXiv preprint arXiv:2412.16504](https://arxiv.org/abs/2412.16504), 2025.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In [Proceedings of the 2016 ACM SIGSAC conference on computer and communications security](https://dl.acm.org/citation.cfm?doid=2969320.2969702), pp. 308–318, 2016.

Noble Ackerson. The Curious Case of ChatGPT’s Privacy Claims, 2024.

ai-robots-txt. ai.robots.txt: A list of ai agents and robots to block. <https://github.com/ai-robots-txt/ai.robots.txt>, 2025. GitHub repository, MIT License.

Anthropic. Can you delete data that I sent via API? Anthropic Privacy Center, 2025a. URL <https://privacy.anthropic.com/en/articles/7996875>.

Anthropic. Is my data used for model training? Anthropic Privacy Center, 2025b. URL <https://privacy.anthropic.com/en/articles/10023580>.

Anthropic. Updates to Consumer Terms and Privacy Policy. Anthropic Privacy Center, 2025c. URL <https://www.anthropic.com/news/updates-to-our-consumer-terms>.

Anthropic. How long do you store my data? Anthropic Privacy Center, 2025d. URL <https://privacy.anthropic.com/en/articles/10023548>.

Yubeen Bae, Minchan Kim, Jaejin Lee, Sangbum Kim, Jaehyung Kim, Yejin Choi, and Niloofar Mireshghallah. Privacy-preserving llm interaction with socratic chain-of-thought reasoning and homomorphically encrypted vector databases. [arXiv preprint arXiv:2506.17336](https://arxiv.org/abs/2506.17336), 2025.

Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. Airgapagent: Protecting privacy-conscious conversational agents. In [Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security](https://dl.acm.org/citation.cfm?doid=3397271.3397303), pp. 3868–3882, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. [arXiv preprint arXiv:2212.08073](https://arxiv.org/abs/2212.08073), 2022.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms. [arXiv preprint arXiv:2309.00667](https://arxiv.org/abs/2309.00667), 2023.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. [arXiv preprint arXiv:2502.17424](https://arxiv.org/abs/2502.17424), 2025.

BGR. ChatGPT privacy: You can opt out of training and keep your history, April 2024. URL <https://bgr.com/tech/chatgpt-got-a-big-privacy-upgrade-heres-whats-new/>.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. [arXiv preprint arXiv:2108.07258](https://arxiv.org/abs/2108.07258), 2021.

Jaydeep Borkar, Matthew Jagielski, Katherine Lee, Niloofar Mireshghallah, David A Smith, and Christopher A Choquette-Choo. Privacy ripple effects from adding or removing personal information in language model training. [arXiv preprint arXiv:2502.15680](https://arxiv.org/abs/2502.15680), 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. [Advances in neural information processing systems](https://dl.acm.org/citation.cfm?doid=3397271.3397303), 33:1877–1901, 2020.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21), pp. 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In International Conference on Learning Representations (ICLR), 2023.

Chaoran Chen, Zhiping Zhang, Bingcan Guo, Shang Ma, Ibrahim Khalilov, Simret A Gebreegzi-abher, Yanfang Ye, Ziang Xiao, Yaxing Yao, Tianshi Li, et al. The obvious invisible threat: Llm-powered gui agents’ vulnerability to fine-print injections. arXiv preprint arXiv:2504.11281, 2025.

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. Copybench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. arXiv preprint arXiv:2407.07087, 2024.

Chinese Social Media Reports. Weibo ai search privacy incident, April 2025.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Szyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. arXiv preprint arXiv:2507.14805, 2025.

Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Liangming Xia, Yijie Bai, Haiqin Weng, and Wenyuan Xu. Sophon: Non-fine-tunable learning to restrain task transferability for pre-trained models. In 2024 IEEE Symposium on Security and Privacy (SP), pp. 2553–2571. IEEE, 2024.

Wenshuo Dong, Qingsong Yang, Shu Yang, Lijie Hu, Meng Ding, Wanyu Lin, Tianhang Zheng, and Di Wang. Understanding and mitigating cross-lingual privacy leakage via language-specific and universal privacy neurons. arXiv preprint arXiv:2506.00759, 2025.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. Reducing privacy risks in online self-disclosures with language models. arXiv preprint arXiv:2311.09538, 2023.

European Data Protection Board. Ai privacy risks and mitigations for large language models. Technical report, EDPB, 2025.

F5 Networks. Rag in the era of llms with 10 million token context windows. Technical report, F5, 2024.

Wei Fan, Haoran Li, Zheye Deng, Weiqi Wang, and Yangqiu Song. Goldcoin: Grounding large language models in privacy laws via contextual integrity theory. arXiv preprint arXiv:2406.11149, 2024.

Fast Company. Google is indexing conversations with chatgpt. <https://www.fastcompany.com/91376687/google-indexing-chatgpt-conversations>, 2025.

Fello AI. ChatGPT Advanced Voice Mode: Everything You Need to Know, September 2024. URL <https://felloai.com/2024/09/chatgpt-advanced-voice-mode>.

Forcepoint Security Labs. Alleged ‘omnigpt’ data breach is a crash course in genai risk. Forcepoint Blog, February 2025.

Vincent Freiberger, Arthur Fleig, and Erik Buchmann. Prisme: A novel llm-powered tool for interactive privacy policy assessment. arXiv preprint arXiv:2501.16033, 2025.

Gadget Review. Grok’s Privacy Disaster: 370,000 AI Conversations Exposed on Google, 2025. URL <https://www.gadgetreview.com/groks-privacy-disaster>.

Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, et al. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. arXiv preprint arXiv:2411.09523, 2024.

General Analysis. Supabase mcp can leak your entire sql database. <https://www.generalanalysis.com/blog/supabase-mcp-blog>, 2025.

GitHub Contributors. Llm osint: Using llms for open source intelligence gathering. https://github.com/sshh12/llm_osint, 2024.

Google. Gemini API Additional Terms of Service. Google AI for Developers, 2025a. URL <https://ai.google.dev/gemini-api/terms>.

Google. Gemini Apps Privacy Hub. Google Support, 2025b. URL <https://support.google.com/gemini/answer/13594961>.

Google. The gemini app can now recall past chats. <https://blog.google/feed/gemini-referencing-past-chats/>, 2025.

Google. How Google handles government requests for user information, 2025. URL <https://policies.google.com/terms/information-requests>.

Google DeepMind. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google, 2025.

Bingcan Guo, Eryue Xu, Zhiping Zhang, and Tianshi Li. Not my agent, not my boundary? elicitation of personal privacy boundaries in ai-delegated information sharing. [arXiv preprint arXiv:2509.21712](https://arxiv.org/abs/2509.21712), 2025a.

Bingcan Guo, Zhiping Zhang, and Tianshi Li. Privi: Assist users in authoring contextual privacy rules with an lm sandbox. In [Proceedings of the 1st ACM Workshop on Human-Centered AI Privacy and Security](https://dl.acm.org/doi/10.1145/3733816.3760756), 2025b. doi: 10.1145/3733816.3760756.

Hacker News Community. Watching o3 guess a photo's location is surreal, dystopian and entertaining. <https://news.ycombinator.com/item?id=43803243>, 2025.

Xinyi Hou, Yanjie Zhao, and Haoyu Wang. On the (in) security of llm app stores. [2025 IEEE Symposium on Security and Privacy \(SP\)](https://ieeexplore.ieee.org/abstract/document/9607400), pp. 317–335, 2025.

Jing Huang, Dify Yang, and Christopher Potts. Demystifying verbatim memorization in large language models. [arXiv preprint arXiv:2407.17817](https://arxiv.org/abs/2407.17817), 2024.

IBM Research. Why larger llm context windows are all the rage. Technical report, IBM, 2024.

Italian Data Protection Authority. Italy Fines OpenAI €15 Million for ChatGPT GDPR Data Privacy Violations, December 2024.

Hanna Kim, Minkyoo Song, Seung Ho Na, Seungwon Shin, and Kimin Lee. When {LLMs} go online: The emerging threat of {Web-Enabled}{LLMs}. In [34th USENIX Security Symposium \(USENIX Security 25\)](https://www.usenix.org/conference/usenixsecurity25/presentation/kim), pp. 1729–1748, 2025.

Isadora Krsek, Anubha Kabra, Yao Dou, Tarek Naous, Laura A. Dabbish, Alan Ritter, Wei Xu, and Sauvik Das. Measuring, modeling, and helping people account for privacy risks in online self-disclosures with ai. [Proc. ACM Hum.-Comput. Interact.](https://dl.acm.org/doi/10.1145/3711029), 9(2), May 2025. doi: 10.1145/3711029. URL <https://doi.org/10.1145/3711029>.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. [Advances in neural information processing systems](https://dl.acm.org/doi/10.1145/3711029), 33: 9459–9474, 2020.

Haoran Li, Wenbin Hu, Huihao Jing, Yulin Chen, Qi Hu, Sirui Han, Tianshu Chu, Peizhao Hu, and Yangqiu Song. Privaci-bench: Evaluating privacy with contextual integrity and legal compliance. [arXiv preprint arXiv:2502.17041](https://arxiv.org/abs/2502.17041), 2025a.

Wenkai Li, Liwen Sun, Zhenxiang Guan, Xuhui Zhou, and Maarten Sap. 1-2-3 check: Enhancing contextual privacy in llm via multi-agent reasoning. [arXiv preprint arXiv:2508.07667](https://arxiv.org/abs/2508.07667), 2025b.

Ryan Little, Lucy Qin, and Mayank Varia. Secure account recovery for a privacy-preserving web service. In *Proceedings of the 33rd USENIX Conference on Security Symposium*, pp. 1993–2010, 2024.

Yupei Liu, Yuqi Jia, Jinyuan Jia, and Neil Zhenqiang Gong. Evaluating {LLM-based} personal information extraction and countermeasures. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 1669–1688, 2025a.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. Protecting privacy in multimodal large language models with mllmu-bench. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4105–4135, 2025b.

Magai. OpenAI’s Court-Ordered Data Retention: What It Means for AI Users, 2025. URL <https://magai.co/openai-court-ordered-data-retention-policy/>.

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. Split-and-denoise: Protect large language model inference with local differential privacy. *arXiv preprint arXiv:2310.09130*, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. Granular privacy control for geolocation with vision language models. *arXiv preprint arXiv:2407.04952*, 2024.

Microsoft. Copilot doesn’t just remember, it also understands you. <https://www.microsoft.com/en-us/microsoft-copilot/for-individuals/do-more-with-ai/general-ai/ai-that-doesnt-just-remember-it-gets-you>, 2025.

Niloofer Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*, 2023.

Niloofer Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint arXiv:2407.11438*, 2024.

Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

NSFOCUS Research. The invisible battlefield behind llm security crisis. Technical report, NSFOCUS Inc., 2025.

OpenAI. How we’re responding to The New York Times’ data demands in order to protect user privacy, 2024. URL <https://openai.com/index/response-to-nyt-data-demands/>.

OpenAI. Sharing feedback, evaluation, and fine-tuning data and api inputs and outputs with openai. <https://help.openai.com/en/articles/10306912>, 2024a.

OpenAI. How we’re responding to the new york times’ data demands in order to protect user privacy. <https://openai.com/index/response-to-nyt-data-demands/>, 2024b.

OpenAI. Data Usage for Consumer Services FAQ. OpenAI Help Center, 2024a. URL <https://help.openai.com/en/articles/7039943>.

OpenAI. Enterprise privacy at OpenAI, 2024b. URL <https://openai.com/enterprise-privacy/>.

OpenAI. How your data is used to improve model performance. OpenAI Help Center, 2024c. URL <https://help.openai.com/en/articles/5722486>.

OpenAI. Privacy Policy, 2024d. URL <https://openai.com/policies/privacy-policy/>.

OpenAI. Temporary Chat FAQ, 2024e.

OpenAI. Introducing operator. <https://openai.com/index/introducing-operator/>, 2025a.

OpenAI. Introducing deep research. <https://openai.com/index/introducing-deep-research/>, 2025b.

OpenRouter.ai. Free models with data logging. <https://openrouter.ai/openrouter/cypher-alpha:free>, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. 2023.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 1314–1331. IEEE, 2020.

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

Badrinath Ramakrishnan and Akshaya Balaji. Assessing and mitigating data memorization risks in fine-tuned large language models. *arXiv preprint arXiv:2508.14062*, 2025.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Reddit User Discussion. Creepy conversations: Scary personal information. https://www.reddit.com/r/ChatGPT/comments/1gox166/creepy_conversations_scary_personal_information/, 2024.

Jaechul Roh, Zachary Novack, Yuefeng Peng, Niloofar Mireshghallah, Taylor Berg-Kirkpatrick, and Amir Houmansadr. Bob’s confetti: Phonetic memorization attacks in music and video generation. *arXiv preprint arXiv:2507.17937*, 2025.

RPC. X suspends personal data training of AI chatbot Grok following Irish DPC pressure, 2024. URL <https://www.rpclegal.com/snapshots/data-protection/autumn-2024/>.

Charlie F Ruan, Yucheng Qin, Xun Zhou, Ruihang Lai, Hongyi Jin, Yixin Dong, Bohan Hou, Meng-Shiun Yu, Yiyan Zhai, Sudeep Agarwal, et al. Webllm: A high-performance in-browser llm inference engine. *arXiv preprint arXiv:2412.15803*, 2024.

Advait Sarkar. Ai could have written this: Birth of a classist slur in knowledge work. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2025.

Kevin Schaul, Szu Yu Chen, and Nitasha Tiku. Inside the secret list of websites that make ai like chatgpt sound smart. *The Washington Post*, April 2023.

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

Xinyue Shen, Yun Shen, Michael Backes, and Yang Zhang. Gptracker: A large-scale measurement of misused gpts. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 336–354. IEEE, 2025.

Silicon Republic. Grok AI is training on user data by default – here’s how to stop it, 2024. URL <https://www.siliconrepublic.com/business/grok-ai-training>.

Li Siyan, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. Papilon: Privacy preservation from internet-based and local language model ensembles. [arXiv preprint arXiv:2410.17127](https://arxiv.org/abs/2410.17127), 2024.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. [arXiv preprint arXiv:2310.07298](https://arxiv.org/abs/2310.07298), 2024.

Surfshark. AI chatbots ranked by data they collect, 2024. URL <https://surfshark.com/research/chart/ai-chatbots-privacy>.

Jingyu Tang, Chaoran Chen, Jiawen Li, Zhiping Zhang, Bingcan Guo, Ibrahim Khalilov, Simret Araya Gebreegziabher, Bingsheng Yao, Dakuo Wang, Yanfang Ye, et al. Dark patterns meet gui agents: Llm agent susceptibility to manipulative interfaces and the role of human oversight. [arXiv preprint arXiv:2509.10723](https://arxiv.org/abs/2509.10723), 2025.

TechCrunch. Here's how to disable x (twitter) from using your data to train its grok ai, July 2024. URL <https://techcrunch.com/2024/07/26/heres-how-to-disable-x-twitter-from-using-your-data-to-train-its-grok-ai/>.

TechCrunch. Elon Musk's X is changing its privacy policy to allow third parties to train AI on your posts, October 2024. URL <https://techcrunch.com/2024/10/17/>.

TechCrunch. Meta fixes bug that could leak users' ai prompts and generated content, July 2025.

TechCrunch. The latest viral chatgpt trend is doing reverse location search from photos. <https://techcrunch.com/2025/04/17/the-latest-viral-chatgpt-trend-is-doing-reverse-location-search-from-photos/>, April 2025a.

TechCrunch. xai adds a memory feature to grok. <https://techcrunch.com/2025/04/16/xai-adds-a-memory-feature-to-grok/>, April 2025b.

Noel Teku, Fengwei Tian, Payel Bhattacharjee, Souradip Chakraborty, Amrit Singh Bedi, and Ravi Tandon. Props: Progressively private self-alignment of large language models. [arXiv preprint arXiv:2508.06783](https://arxiv.org/abs/2508.06783), 2025.

Theori Research. Deepseek security, privacy, and governance: Hidden risks in open-source ai. Technical report, Theori Inc., January 2025.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In [25th USENIX security symposium \(USENIX Security 16\)](https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer), pp. 601–618, 2016.

Truffle Security. 12,000 secrets found in common crawl training data. Technical report, Truffle Security Co., January 2025.

U.S. District Court. The new york times company v. microsoft corporation and openai inc., 2024. Case No. 1:23-cv-11195.

U.S. District Court. Bartz et al. v. anthropic pbc, 2025. Judge William Alsup ruling.

Yash Vekaria, Aurelio Loris Canino, Jonathan Levitsky, Alex Ciechonski, Patricia Callejo, Anna Maria Mandalari, and Zubair Shafiq. Big help or big brother? auditing tracking, profiling, and personalization in generative ai assistants. [arXiv preprint arXiv:2503.16586](https://arxiv.org/abs/2503.16586), 2025.

VentureBeat. Sam altman calls for 'ai privilege' as openai clarifies court order to retain temporary and deleted chatgpt sessions. <https://venturebeat.com/ai/sam-altman-calls-for-ai-privilege/>, January 2025.

Wald. ChatGPT Data Leaks and Security Incidents (2023-2025): A Comprehensive Overview, 2025. URL <https://wald.ai/blog/chatgpt-data-leaks-and-security-incidents>.

Shenao Wang, Yanjie Zhao, Xinyi Hou, and Haoyu Wang. Large language model supply chain: A research agenda. *ACM Transactions on Software Engineering and Methodology*, 34(5):1–46, 2025.

Weibo. Weibo daily active user statistics, 2025.

Simon Willison. The lethal trifecta. <https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/>, June 2025.

xAI. Enterprise FAQs, 2025a. URL <https://x.ai/legal/faq-enterprise>.

xAI. Privacy Policy, 2025b. URL <https://x.ai/legal/privacy-policy>.

Xiaohongshu User Discussion. Chatgpt voice mode privacy concerns, 2024.

Xiao Zhan, Juan Carlos Carrillo, William Seymour, and Jose Such. Malicious llm-based conversational ai makes users reveal personal information. *arXiv preprint arXiv:2506.11680*, 2025.

Shuning Zhang, Lyumanshan Ye, Xin Yi, Jingyu Tang, Bo Shui, Haobin Xing, Pengfei Liu, and Hewu Li. "ghost of the past": identifying and resolving privacy leakage from llm's memory through proactive user interaction. *arXiv preprint arXiv:2410.14931*, 2024a.

Zhiping Zhang, Bingcan Guo, and Tianshi Li. Privacy leakage overshadowed by views of ai: A study on human oversight of privacy in language model agent. *arXiv preprint arXiv:2411.01344*, 2024b.

Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 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*, pp. 1–26, 2024c.

Zhiping Zhang, Chenxinran Shen, Bingsheng Yao, Dakuo Wang, and Tianshi Li. Secret use of large language model (llm). *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–26, 2025a.

Zhiping Zhang, Yi Evie Zhang, Freda Shi, and Tianshi Li. Autonomy matters: A study on personalization-privacy dilemma in llm agents. *arXiv preprint*, 2025b.

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

Jijie Zhou, Niloofar Mireshghallah, and Tianshi Li. Operationalizing data minimization for privacy-preserving llm prompting. *arXiv preprint*, 2025a.

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. Rescriber: Smaller-llm-powered user-led data minimization for llm-based chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–28, 2025b.

Zira Daily. Hackers can easily read what you say to ChatGPT, other AI services, 2025. URL <https://ziradaily.com/news/88713>.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium*. USENIX Association, 2025.