Understanding Users' Privacy Perceptions Towards LLM's RAG-based Memory

Shuning Zhang
zsn23@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Rongjun Ma
rongjun.ma@aalto.fi
Aalto University
Espoo, Finland

Ying Ma
ying.ma1@student.unimelb.edu.au
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia

Shixuan Li
li-sx24@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Yiqun Xu
xuyiqun22@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Xin Yi*
yixin@tsinghua.edu.cn
Tsinghua University
Beijing, China

Hewu Li
lihewu@cernet.edu.cn
Tsinghua University
Beijing, China

Abstract

Large Language Models (LLMs) are increasingly integrating memory functionalities to provide personalized and context-aware interactions. However, user understanding, practices and expectations regarding these memory systems are not yet well understood. This paper presents a thematic analysis of semi-structured interviews with 18 users to explore their mental models of LLM's Retrieval Augmented Generation (RAG)-based memory, current usage practices, perceived benefits and drawbacks, privacy concerns and expectations for future memory systems. Our findings reveal diverse and often incomplete mental models of how memory operates. While users appreciate the potential for enhanced personalization and efficiency, significant concerns exist regarding privacy, control and the accuracy of remembered information. Users express a desire for granular control over memory generation, management, usage and updating, including clear mechanisms for reviewing, editing, deleting and categorizing memories, as well as transparent insight into how memories and inferred information are used. We discuss design implications for creating more user-centric, transparent, and trustworthy LLM memory systems.

CCS Concepts

• **Security and privacy → Usability in security and privacy**.

*Corresponding author.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Keywords

Memory, Large Language Model, Privacy Perception, Personalization, Trade-offs

ACM Reference Format:
Shuning Zhang, Rongjun Ma, Ying Ma, Shixuan Li, Yiqun Xu, Xin Yi, and Hewu Li. 2018. Understanding Users' Privacy Perceptions Towards LLM's RAG-based Memory. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Large Language Models (LLMs) like ChatGPT[1], Gemini[2], and Kimi[3] are rapidly evolving from stateless tools into personalized assistants. A key technology driving this shift is the integration of memory, which allows LLMs to retain information across conversations to provide more coherent and contextually aware interactions. While this promises enhanced utility, it also introduces a significant and complex privacy challenge. By creating a persistent record of user interactions, these systems build detailed profiles that can include sensitive thoughts, personal plans, and confidential information, moving beyond transient queries to continuous user data collection.

This capability creates a fundamental tension between personalization and privacy of LLM's memory. On one hand, users desire the efficiency and tailored responses that memory enables. On the other, the opacity of how these systems remember, analyze and utilize personal data can undermine a user's sense of control and informational self-determination. The "black box" nature of LLMs exacerbates this issue, leaving users unable to fully understand the scope of data being collected in memories or the associated privacy risks. Specifically, for RAG-based memories, they were usually extracted from users' dialogues, as per ChatGPT, and used for

[1]https://chat.openai.com/
[2]https://gemini.google.com/
[3]kimi.moonshot.cn

latter personalized dialogues as context information. It may cause potential risks if those RAG-based memories are utilized for training or disclosed to untrusted third-party services. However, how users understand the RAG-based memories, what's their practice and challenges remains under-explored. This study therefore investigates how users perceive RAG-based memories, manage the conflict between privacy and utility, and what's their challenges. We aim to first understand users' mental models of memories as their mental model largely influence whether they could are aware of the potential privacy risks and could effectively control. We then identified their current calculus around the privacy-utility trade-offs and the protective strategies, as private memories inherently possess the conflicts between personalization and privacy leakage. We finally understand users' challenges and expectations for improving RAG-based memory systems. Specifically, we seek to answer the following questions:

RQ1. What is users' mentral models of RAG-based LLM memories?

RQ2. How do users navigate the trade-offs between utility and privacy when using or considering LLM memory, and what privacy protection strategies do they currently employ?

RQ3. What are the challenges users face when attempting to align an LLM's memory behavior with their privacy goals?

We conducted semi-structured interviews with 18 Chinese participants from diverse backgrounds with varying levels of experience. To anchor the discussion in a real-world context and elicit concrete responses, the interview was centered on ChatGPT's memory feature as a prominent case study.

Towards RQ1, we find four prominent mental models of users: users regard memory as a transient, dialogue-specific buffer, as an extension of the core training data, as an active information processing mechanism, and some users explicitly acknowledged the lack of understanding. These mental models reflected different understandings and cognitive processing logics of LLMs' memories. Towards RQ2, our research shows that users are not passive but are active agents performing a continuous privacy calculus, weighing functional benefits against perceived risks. The primary benefits involved enhance personalization and efficiency, improvec continuity and reduced redundancy, and LLM to be a better companion of assistant. The primary drawbacks include problems such as confidentiality and data leakage, profiling and unwanted persuasion, aggregation and re-identification, unauthorized secondary use, and the lack of epistemological certainty. Users employ proactive protective strategies, such as strategic privacy disclosure, proactive input obfuscation and refusal to use or workarounds. Towards RQ3, we identify an unequivocal user mandate for a fundamental shift toward user-directed memory systems built on granular control and transparency across the entire data lifecycle. This includes explicit consent at the point of generation, comprehensive management interfaces for editing and deletion, purpose limitation controls during usage, and direct agency over system-inferred information. To sum up, our contributions are:

- We characterize users' diverse mental models of LLM memory and their privacy risk perceptions.
- We investigate the privacy calculus and protective strategies users employ, such as data minimization and anonymization.
- We identify the critical privacy challenges user face and distill our findings into design implications for user-centric, privacy-preserving memory systems.

2 Related Work

2.1 Privacy Protection of Text-based LLMs

Research on privacy protection for LLMs follows two primary lines: privacy risk evaluation and direct defense mechanisms. Risk evaluation, which underpins any protective effort, is supported by several toolkits and benchmarks. LLM-PBE [17] supplies a structured framework for assessing risks throughout the model lifecycle, and PrivLM-Bench [16] offers a standardized benchmark to quantify data leakage. Specialized instruments such as PrivacyLens [30] evaluate compliance with privacy norms, while ProPILE [14] probes models to detect possible PII exposure.

Direct defense mechanisms are engineered to intervene at different points. Some operate at the data interface: OneShield [2] filters both user inputs and model outputs, while Rescriber [50] leverages an LLM to minimize sensitive content in queries in real time. In contrast, other strategies intervene at the model level. The CPPLM paradigm [35], for example, embeds safeguards directly into the fine-tuning process to protect inference-time privacy.

2.2 Understanding Users' Privacy Concerns of LLMs

Research into users' privacy concerns in digital contexts began early, showing that users often self-regulate by selectively withholding information or avoiding services perceived as high-risk [5, 24, 26]. In the LLM domain, this manifests as a persistent trade-off between privacy, utility, and convenience [47], frequently giving rise to a "privacy paradox" in which users tolerate greater information leakage in exchange for higher utility [46]. Similarly, users of conversational agents tend to exhibit fewer privacy concerns than non-users, sharing sensitive lifestyle and health data while withholding direct identifiers [51].

These behaviors are further amplified by cognitive biases and system design. Users often rely on inaccurate mental models of LLM data flows and are vulnerable to dark patterns, which impedes their ability to grasp real privacy risks [22, 47]. Additionally, the anthropomorphic presentation of LLMs encourages oversharing, as users overestimate the system's capabilities and attribute human-like understanding to it [11, 33]. This tendency heightens privacy exposure and can be exploited for malicious purposes [13, 28].

2.3 Memory in LLM-driven Agents

The capacity to maintain conversational history is a foundational feature of LLM-driven agents, enabling coherent and contextually aware human–AI interaction [10, 20]. Early methods appended the full chat history to the model context. However, as conversations lengthened, agent performance degraded due to distraction and information loss—often called "lost in the middle" [21, 31]. Consequently, research shifted toward more sophisticated memory architectures. These include recursive summarization and refinement techniques to distill salient information and reduce redundancy [12, 32, 37, 38, 49], as well as selective retrieval-based systems

that store memories in unstructured or layered repositories and fetch them based on conversational relevance [3, 27, 34, 41, 48].

Despite these architectural advances, memory storage and retrieval remain opaque to end users, impeding effective task integration. Recent work has explored interactive memory systems that give users direct control via operational "sandboxes" for manual editing [12] or visual interfaces for intuitive organization [40]. Zhang et al. [44] further examined user challenges and developed techniques to support memory use. However, these studies have not conducted an in-depth analysis of users' mental models of memory systems nor explicitly examined the core trade-off between personalization and privacy, which this work addresses.

3 Methodology

The study employed a qualitative approach to investigate user perceptions and expectations of memory functionalities in LLMs. We utilized semi-structured interviews to gather rich, in-depth data from participants.

3.1 Participant Recruitment and Demographics

We recruited 18 participants for this study through distributing questionnaires online. The participants represented a range of academic and professional backgrounds, including fields such as engineering, life sciences, social sciences, IT and design. Table 1 showed the demographics of participants. The study was approved by the Institutional Review Board (IRB) of our institution, and each participant was compensated 100 RMB for their time.

3.2 Interview Design and Procedure

We conducted semi-structured interviews in Chinese, which allow for flexibility in exploring emergent themes while ensuring that core topics related to LLM memory were covered with each participant. The interviews focused on participants' mental models of LLM memory, their current practices and experiences, perceived advantages and disadvantages, willingness to share different types of information, privacy concerns and their expectations of ideal LLM memory systems across its lifecycle (generation, management, usage and updating). All interviews are conducted via Tencent meeting online[4], and we recorded and transcribed these for analysis.

3.3 Data Analysis

The interview data, comprising qualitative notes and direct participant quotes, was analyzed using thematic analysis. This involved an iterative process wherein two researchers first familiarized themselves with the data. Initial codes were then inductively generated based on participants' discussions of memory mechanisms, benefits, drawbacks, privacy risks, and desired features. Following this, a collaborative process was undertaken to discuss these initial codes and develop a unified codebook. To ensure consistency, two researchers then independently coded 20% of the data, and inter-rater reliability achieved Cohen's Kappa of 0.90. They then independently coded the rest of the dataset. Subsequently, these codes were collated into potential themes and sub-themes, which were then reviewed, defined, and refined through further discussion to ensure they accurately captured the nuances of participants' experiences and perspectives. The analysis specifically focused on identifying users' mental models of LLM memories' privacy risks, their current mitigation practices and encountered challenges, and their expectations for the design of privacy-preserving LLM memory systems across the lifecycle.

[4]https://meeting.tencent.com/

4 Results

We started from delineating users' mental models (RQ1), especially around privacy risks of memories. We then detailed their practices, especially concerning how they trade privacy for utility in utilizing memories (RQ2). Finally, we outlined their challenges and future expectations (RQ3).

4.1 RQ1: User Mental Models towards LLM Memory

Participants' conceptualization of how LLM memory functions were markedly varied and often informed by inaccurate analogies. The analysis revealed four dominant themes in their mental models: (1) memory as a transient, dialogue-specific buffer, (2) memory as an extension of core training data, (3) memory as an active information processing mechanism, and (4) a widespread acknowledged lack of understanding.

Mental Model 1: memory as a transient, dialogue-specific buffer.

A substantial group of users conceptualized LLM memory as ephemeral, existing only within the confines of a single, continuous interaction or dialogue session. This model posits that any contextual understanding gained by the LLM is reset once a conversation window is closed, preventing memory from persisting across separate interactions. A primary misconception articulated by P1 was the belief that *"the LLM's memory can only be maintained in the same dialogue window."* This view was explicitly shared by P5, who stated that memory *"exists in the same dialogue, but not across dialogues."* This mental model has direct behavioral consequences, such as users starting new chats to "reset" the LLM's context when it becomes stuck on a flawed instruction.

Mental Model 2: memory as an extension of core training data.

Another prevalent mental model treats LLM memory as a mechanism that directly augments the model's foundational training dataset. Participants with this view envision their conversations being absorbed into the LLM's public knowledge base, akin to a continuous training process, potentially compromising their privacy. This model blurs the line between a private, personalized memory and the public, generalized knowledge of the AI. P3 wondered if memory was achieved *"through training on massive data,"* considering the feature a *"long-term memory model ... built upon the original training data. ... comprehensively collecting new information ... then conducting further training."*

This perspective was clearly echoed by P16, who questioned if memory involved *"taking past chat data, re-labeling it, and then put[ting] it into its model for training"*. This conceptualization often leads to heightened privacy concerns, as it implies that personal or proprietary information could be permanently integrated into the model and potentially exposed to other users. P16 later articulated

Table 1: Demographics of users. ('Par' denotes participant ID.)

Par	Age	Highest education	Usage experience	Occupation /
P1	23	Ph.D.	ChatGPT, kimi, openai chat sider, dxyz, mobile poe, slack	Clean combustion
P2	21	High school	ChatGPT, Gemini, Kimi	Industrial engineering
P3	22	Bachelor	ChatGPT3.5/4o	Naval architecture and ocean engineering
P4	21	Bachelor	ChatGPT, ChatGPT4.0	Engineering mechanics
P5	24	Master	ChatGPT, Wenxinyiyan, Kimi, Bing	Other
P6	20	Bachelor	ChatGPT	Electronic packaging
P7	19	High school	GPT, Kimi	Electronic science and technology, AI research
P8	24	Bachelor	Kimi, ChatGPT3.5	Life science
P9	22	Master	ChatGPT3.5, Kimi, Wenxinyiyan	Social science
P10	26	Master	ChatGPT, ChatGPT4.0	Civil engineering, rock mechanics
P11	24	Master	ChatGPT, Tongyi, PI, free, no payment	Other
P12	20	Bachelor	ChatGPT, Gemini	IT related, software engineering
P13	24	Master	ChatGPT, Kimi	Not disclosed
P14	27	Master	GPT, Wenxinyiyan, Kimi, Poe, TongYi, Doubao, Xiaomei	Design related
P15	23	Master	ChatGPT, ChatGLM, Kimi, Doubao	Electronic information, digital media
P16	24	Master	ChatGPT, Kimi, Doubao	Design studies
P17	22	Bachelor	ChatGPT, Midjourney, Stable Diffusion	Design profession
P18	24	Master	ChatGPT	Software engineering

this fear, stating, *"I would be worried ... that this data is being used for training from the very beginning"*. This model suggests a permanent, irreversible form of memory, fundamentally altering the core model rather than creating a separate, user-specific memory layer.

Mental Model 3: memory as an active information processing mechanism.

A third group of participants, often those with technical exposure, described a dynamic mental model in which memory is the result of an active information-handling process. This perspective moves beyond simple storage, suggesting the LLM intelligently curates and structures memory content along a spectrum of perceived complexity. At the simpler end of this spectrum, some users envisioned a process of contextual accumulation. P15 for example, described the mechanism as one that *"is to accumulate the information in your dialogue into the original question each time ... It is the stacking of text content."* More sophisticated models involved active summarization and selective extraction. P12 proposed that *"the LLM's memory is its own summary of the input."* This concept was most elaborately detailed by P18, who envisioned a discerning agent that actively decides what to remember by conducting its own *"analysis or summary"* to *"extract what points it needs to remember ... and place these extracted contents ... in a separate 'memory' "*. P18 further speculated that this process could be user-directed, for instance, when a prompt explicitly instructs the model to *"remember something"*. This theme also encompassed a high-level, albeit incomplete, awareness of the underlying technology, with P6, for instance, noting that the process *"relies on neural networks and training with large parameters"* while admitting to not being *"very clear on the specifics"*.

Mental Model 4: acknowledged lack of understanding.

Across all participant groups, there was a significant and openly acknowledged lack of a clear or confident mental model for LLM memory. This uncertainty persisted regardless of the user's technical background or frequency of use, highlighting the *"black box"* nature of the memory.

Many participants were direct about their confusion. P9 stated plainly, *"I don't know how it works"*, and P11 was *"unclear if there is a memory function"* at all. Some users held definite, but incorrect, ideas born from this uncertainty. For example, P8 misunderstood the feature as one where *"you can upload a document and then it can output content based on user preferences"*. Even P7, an AI researcher, confessed to *"not having used the memory function"* and thus lacked an experiential basis for understanding it. This gap was also present in users who had attempted to learn more. P6, despite referencing neural networks, qualified his explanation by stating, *"I'm not very clear on the specifics"*.

4.2 RQ2: The Privacy Calculus: Navigating Trade-Offs and Employing Protective Strategies

Participants described both benefits and concerns related to LLM memory, reflecting a privacy calculus. This section outlines the perceived benefits (section 4.2.1) and concerns (section 4.2.2) that shaped how they evaluated this trade-off.

4.2.1 Perceived benefits of LLM memory. Participants identified key advantages that motivate data sharing, involving increased personalization, improved interaction efficiency, continuity across tasks and the potential for symbiotic user-LLM relationship.

Enhanced personalization and efficiency: The most frequently cited benefit was the potential for LLMs to deliver personalized and efficient interactions by remembering user preferences and context. P1 anticipated that memory would lead to *"Personalization, asking questions in a certain field will yield answers more aligned with what*

is desired ... saves some time." This sentiment was echoed by P2, who highlighted *"more efficient dialogue, providing more valuable results"* and P18, who envisioned that memory would make the LLM *"be more personalized, it aligns better with my usual habits ... it will understand me better, sometimes I don't need to explicitly state what information I need."* P9 also noted the benefit of saving time by enabling *"refined input of one's own needs, outputting answers"*. Furthermore, P10 appreciated how memory could *"help me obtain standardized answers, layer by layer, following my logic."* P8 elaborated on this, expecting memory to lead to explicit understanding, adherence to instructions, avoidance of misinterpretation, and tailored recommendations, ultimately making information access convenient.

Improved continuity and reduced redundancy: Users valued memory for its ability to maintain context over time, reducing the need for repetitive input, especially for ongoing or complex tasks. P5 appreciated that *"When working on the same large assignment, I don't have to input everything repeatedly"*, a point also made by P11. P15 found it useful that *"I can omit some key domain definitions asked in the first question ... it simplifies the complexity of my questions"*. P7 also saw convenience in the LLM remembering basic information when initiating new dialogues.

LLM as a better companion of assistant: Some participants saw memory as a way to foster a relational interaction with the LLM. P7 suggested it would be useful if the LLM could *"remember what was confided"* when used as a confident. P15 articulated a desire for a familiar interaction, stating, *"it would be like someone who knows you ... there would be a sense of closeness."* Our participants shared that this sense of companionship often outweighed concerns about privacy.

4.2.2 Concerns of LLM memory. Counterbalancing these benefits is a wide spectrum of privacy threats that users are concerned about.

Confidentiality and data leakage: Users expressed significant concern about the unauthorized disclosure of sensitive information. This included the leakage of unpublished professional work (P1), proprietary source code (P3), and financial details like bank statements (P17).

Profiling and unwanted persuasion: A primary fear was that remembered information would be used to create detailed user profiles for malicious or commercial purposes. P6 identified the risk of *"exposure of personal habits and preferences, leading to targeted information, popular scams, ad calls"*. P7 worried about the LLM *"mastering private life activities, homework and work content, research, life and work."*

Aggregation and re-identification: Users with technical backgrounds feared the power of data aggregation. P12 worried that combining fragmented pieces of personal information could lead to a *"comprehensive profile"* an that *"cross-verification leading to inference"* could re-identify an individual from seemingly innocuous data points like a school and student ID.

Unauthorized secondary use: The concern that their conversational data might be used for other purposes, such as modeling training, without their full understanding was a key issue for users like P16.

Lack of epistemological certainty: A profound concern was the skepticism about whether user actions, such as deletion, had any real effect. P18 expressed a deep-seated distrust, stating, *"Although I can delete this memory on the client or web end ... I remain skeptical whether it will be deleted from their database."* This uncertainty undermines the perceived effectiveness of any user-facing privacy controls.

4.2.3 User-Devised Protective Strategies. In response to this calculus, users employ a range of protective strategies, moving from content curation to outright rejection of the technology.

Strategic privacy disclosure: A primary strategy is the active management of privacy categories disclosed. Users curate the information they share, creating clear distinctions between permissible and forbidden data. Users are generally amenable to LLMs remembering non-sensitive, task-oriented information that provides benefit for their future tasks. This includes professional context like their *"writing documents' customs"* (P3), academic materials like *"coding formats"* and *"textbook key points"* (P6), and project-specific data like *"interview transcripts"* for summarization (P11). Users establish private zones for sensitive information. This includes personally identifiable information (PII) and financial data ("ID numbers, contact information, Alipay, bank card numbers" (P6)), unpublished intellectual property ("core research projects" (P4), "unpublished papers" (P10)), and sensitive personal data such as political views (P9) or a permanent home address (P8). P10 articulated a sophisticated desire for selective processing, hoping the model would *"remember needed knowledge, but not provided habits, personal ways of speaking"*.

Proactive input obfuscation: When users choose to disclose the data, many engage in proactive data minimization and obfuscation. This includes filtering out sensitive details, with P5 admitting: *"I filter out significant personal data without entering."* It also involves anonymization, such as P12's practice of *"directly erasing information that needs to be anonymized"*.

Refusal to use, or workarounds: When the perceived privacy risk is too high or the system is deemed untrustworthy, users' strategies resort to rejection or the creation of workarounds. Some users, like P12, explicitly reject the feature, stating, *"I don't want this kind of memory ... I hope the LLM's executions are mutually independent."* Others develop practical workarounds to bypass flawed memory systems, such as P17's decision to *"start a completely new environment"* to escape an inaccurate memory loop. Even the preference to *"habitually start from scratch"* (P17) can be seen as a protective strategy to ensure data accuracy and avoid the risks of a faulty memory system.

4.3 RQ3: Challenges in Aligning LLM Memory with User Privacy Goals

To answer RQ3, our analysis identified critical challenges user face when attempting to align LLM's memory behavior with their privacy goals. These challenges are experienced by users as concrete frustrations and unmet needs, manifesting across the entire memory lifecycle, from the moment a memory is created to how it is used.

4.3.1 Inaccurate Memories. A foundational challenge for users is the inaccuracy of LLM's memories, that can remember information inaccurately, apply it in the wrong context, or update it in an uncontrolled manner, making it difficult to build a stable and reliable personalized experience. This challenge is evident in user reports of flawed recall. P17 described a frustrating experience where the LLM *"seemed to always remember my first requirement and kept modifying the code according to the first requirement's standard,"* forcing them to *"start a completely new environment"* to escape the faulty memory. Users report that the system might *"remember things incorrectly"* or *"associate an answer with a different question"* (P1), and uncritically memorize *"human-inputted text [that] might have errors"* (P2). In particular, P1 noted that *"questions related to combustion asked in the past would later appear in unrelated course extension content,"* a contextual error that undermines the memory's utility. The challenge is compounded during memory updates, where users fear the system *"cannot guarantee the authenticity of updated data, covering previous things."* (P6)

Users' articulated needs reveal their struggle to overcome this challenge. They desire agency over updates, wanting to be notified of *"what memory was replaced"* (P5) or to have a say when conflicts arise, for instance, through *"a warning icon [that] appears ... shows me the update, [and] asks if I accept"* (P8). The desire for *"a correction mechanism"* to fix errors after the fact (P11) further underscores the core challenge of maintaining an accurate and trustworthy memory record.

4.3.2 Lack of Meaningful Control and Transparency. Another challenge is the lack of meaningful user control and transparency across the memory lifecycle. Users consistently described feeling powerless and uninformed about how the system operates, which directly prevents them from aligning its behavior with their privacy goals.

The memory creation process is opaque and lacks user agency. This lack of intelligent filtering is a key challenge. As P6 argue, users *"can't choose to improve or modify memories,"* preventing true personalization. To overcome this, users demand direct control over what is committed to memory. This includes the need to *"manually control what is added and what is not"* (P2), and the ability to *"decide whether to add after generation"* (P9). Some users, like P13, feel challenges by the system's over-summarization and argue that memory *"should be user-edited, no need for the LLM to refine."*

Regarding management, users also think memory management tools are rudimentary, with some even unknown of such tools. This leads to confusion, as expressed by P10 who *"just enabled the function, and never opened the management interface,"*. In contrast, users desire sophisticated organizational tools, envisioning a memory structured like *"a book with a table of content"* (P1) with *"automatic classification, automatic clustering, tree structure"* (P12), and even context-specific *"memory module classification"* (P4). A critical need is for fine-grained control over individual memories, including *"information filtering, viewing and editing"* (P3) and simple, direct deletion (P5,P9). Furthermore, users are challenged by the lack of temporal control and desire features like a *"timed deletion function"* (P6) or the ability to separate *"recent memory and long-term memory"*. (P1)

Regarding usage, the current usage has no options for users to control, and is also non-transparent. P10 articulated this frustration, stating the system *"will only tell you what the updated memory is, won't tell you which memory was used. Maybe it defaults to use all memory?"* This makes it troublesome for users to manage the context of their interactions (P9). To align the system's usage with their goals, users expect numerous controls. They desire the ability to selectively activate memories for specific tasks, wanting to *"distinguish, to use this part of the memory and not other parts, distinguish personas"* (P4). This led to users imagining features like a *"browser-like incognito window"* to temporarily disable memory (P8) or the ability to *"switch identities"* between different memory contexts (P11). To overcome the challenge of opacity, users demand transparency in how memory influences responses. P7 asserted that the LLM *"has to know why it used this memory,"* and P18 found such displays "necessary" to reduce the anxiety of interacting with a "black box".

4.3.3 Challenge to Manage Opaque and Uncontrolled Inferences. Perhaps the most complex challenge users face is in managing information they never explicitly provided but that the LLM has inferred. Users are aware of, but has no countermeasures against this capability, which they perceive as a significant privacy threat.

Users' challenges stem from the opacity of the inference process. Users worry about *"additional privacy risks"* like their physical location being inferred and tracked (P1, P9) and are unsettled by the prospect of constant, daily reasoning about their live habits (P7). P6 noted with certainty that the LLM can infer his profession from his queries. This leads to a feeling of being profiled and judged, and P11 also described this inference is like *"running naked"*.

To overcome this challenge of opaque inference, users demand radical transparency and control. They want to be explicitly told *"what is inferred based on the information"* and have the power to *"delete it"* (P8). They assert a right to know *"if explicit private information... was inferred"* and even see the *"confidence or accuracy"* of that inference (P9). The desired controls are equally robust, ranging from tools to *"blur some personal characteristics"* (P2), to commands that tell the LLM to *"shut up"* about certain topics (P7), to high-level policy interventions that *"strengthen supervision and regulation"* (P4). This reveals that for users, aligning the system's behavior with their goals is not just about managing what they input, but also about governing what the system creates on its own.

5 Discussions and Future Work

Our study reveals a significant disconnect between the functionality of emerging LLM memory systems and the mental models, expectations, and concerns of users. The findings highlight a critical need for more transparent, controllable, and user-centric memory designs. In this section, we discuss the core tensions emerging from our data, the discrepancy between user mental models and system reality, and the perennial challenge of balancing personalization with privacy, before outlining concrete design implications and directions for future work.

5.1 Memory and the Privacy Implications

There are different types of memories, and even during the interview, some memories participants mentioned are not RAG-based memories. Participants were found to confuse different types of memories' functionalities, such as believing that ChatGPT only has

memories that directly uses the past dialogue. This phenomenon is constantly evolving in the current AI agent age, as more and more agents have different types of memories, such as contextual memories, RAG-based memories, epistemological memories, etc.

While these memory types share common privacy concerns—such as data persistence, inference risks, and lack of user control—they differ significantly in their implementation and privacy implications. Contextual memories typically maintain conversation history within session boundaries and are often perceived as more transient, aligning with participants' Mental Model 1 of dialogue-specific buffers. In contrast, RAG-based memories involve explicit extraction and storage of user information across sessions, creating persistent user profiles that can be retrieved and applied in future interactions. This persistence amplifies privacy risks through potential data aggregation and cross-session inference, as evidenced by participants' concerns about "comprehensive profiling" and "cross-verification leading to inference" (P12).

Epistemological memories, which store factual knowledge and learned concepts, present different challenges as they blur the line between personal data and general knowledge, potentially leading to the privacy risks associated with Mental Model 2 where participants feared their data being integrated into training datasets. The confusion among participants regarding these distinctions has important implications for privacy risk assessment, as users operating under incorrect mental models may apply inappropriate privacy protection strategies. Although our work only sheds light on RAG-based memories, we envision that future work could solve problems around contextual memories, or other types of memories, by developing differentiated privacy controls and transparent communication about each memory type's specific characteristics and associated privacy implications.

The integration of memory into LLMs introduces a complex landscape of privacy considerations, largely shaped by the significant disconnect between users' mental models and the system's actual operations. While not all information retained by an LLM constitutes a privacy risk, the opacity of these memory systems creates potential vulnerabilities. This research, as outlined in RQ1, reveals that users' varying and often inaccurate conceptualization of how memory functions can directly exacerbate these risks.

One prevalent mental model, which conceives of memory as a transient, dialogue specific buffer, may foster a false sense of security. Users operating under the assumption that all contextual information is purged at the end of a session are more likely to disclose sensitive data, believing it to be ephemeral. This misconception significantly increases the risk of inadvertent data exposure, as the system may retain and utilize this information in ways the user neither anticipates nor consents to.

Conversely, the mental model of memory as an active information processing mechanism introduces a different set of privacy challenges. While this model aligns more closely with the sophisticated capabilities of advanced AI, it elevates the risk of inferential privacy breaches. The system's perceived autonomy to analyze, summarize and draw conclusion from user inputs means that sensitive attributes, such as health status, political affiliation, or personal habits, can be inferred without ever being explicitly stated by the user. This capability can create a chilling effect, fostering a sense of being constantly monitored that may lead to user self-censorship and a consequent erosion of free expression and autonomy.

Finally, the widespread lack of understanding among users highlights a fundamental gap in system transparency and user education. When users are unable to form accurate mental models, their ability to provide informed consent and exercise meaningful control over their personal data is fundamentally compromised. This ambiguity forces users to rely on folk theories or inaccurate analogies, impeding the adoption of privacy-preserving behaviors and undermining trust in the technology. Effectively addressing the privacy implications of LLM memory, therefore, requires not only robust technical safeguards but also a concerted effort to provide clear, accessible explanations of how these complex systems operate.

5.2 Trade-offs in Memory Management and Usage

Current memory management systems often operate proactively, requiring minimal user intervention. Alternatively, providing users with more explicit choices and consent can enhance their agency and control [23, 25]. This approach, however, introduces a known trade-off: increasing user control can also increase cognitive load and privacy fatigue [8], potentially diminishing the user's sense of agency [42, 43]. To effectively operationalize memory based on user expectations, systems must first understand users' nuanced privacy preferences [1], using both implicit and explicit methods [39]. Based on this understanding, a system can adjust its degree of proactivity, offering simple controls for key decisions while implementing other protections through methods like privacy by design [7].

Users' privacy challenges with LLM memory align with the traditional privacy calculus model [15]. Even in OpenAI's ecosystems, users could emphasize privacy by using the "Incognito mode", although it has no personalization features, and acts as an extrema of this balancing. Memory augments conversation with long-term context, enabling a higher degree of personalization than traditional recommendation systems [1]. This capability, in turn, introduces a wider range of privacy preferences that require consideration. This calls for new interaction designs tailored to memory systems that can facilitate user preference selection [40]. Unlike approaches that focus on anonymizing discrete text inputs [45], managing the privacy-personalization trade-off for persistent memory requires first communicating these compromises to the user to align their mental models. After establishing this alignment, the system can more accurately collect and model user feedback [39], even within the ambiguous context of text-based interactions.

Users' perception towards memory reflects their willingness of stronger agency on their memory control, echoing the long-discussed balancing between users' and systems' agency [19, 42]. Although users' responses in our study also varies, with some desiring granular supervision, others with coarsed ones, their consensus is that the current control is far from enough.

5.3 Cultural Nuances of Memories

Participants' usage around memory primarily centered around a balance between personalization and privacy risks, which reflect their privacy calculus [15]. The privacy risks and preferences may

subject to culture nuances, as reflected in the prior work [36]. For example, Xu et al. [36] found annotators from an Eastern country like Japan paid less attention to exposing their individual preferences. The difference in memory's privacy preferences may be subject to the power distance [9] and culture norms [18]. Besides, different cultures may involve different valuing of the memory's balance. with some culture valuing personalization more and others valuing privacy more [29]. Therefore, for our results to be generalizable to Western cultures, guided by prior work [36], we hypothesized that a Western country may be more cautious along the privacy-utility balance. We also regard the detailed examination of the cultural nuances as our future work.

5.4 Limitations

We acknowledged that this paper has two limitaitons. As our study is certered on Chinese users under Chinese regulations, there may be regulatory nuances and difference (e.g., GDPR-applied regions or CCPA-related regions). Our participants are also biased towards young students, which possess higher education and literacy than the average. As we find they are subject ot privacy risks, we believe more efforts is needed to investigate and prevent the privacy risks associated with memories for the generic public. We acknowledged that different cultures has nuanced privacy preferences differences and regard the trade-offs' examination in other cultures and regions as our future work. Besides, we primarily target ChatGPT, which is the product most participants has used which has memory features. There are other products like Gemini or open-sourced agent frameworks which may have different implementations of memory features. We regarded them as the future work.

5.5 Design Implications

Our findings call for designs that address three interdependent aspects of the user-system relationship: the systems' architecture, its communicative interface, and the dynamics of its interaction:

The architectual layer: designing contextual aware memories

The current memory system collapses from the diverse users' context to a single data stream, which requires a shift from a monolithic memory to a modular, context-aware architecture. Systems should be architected around distinct memory "workspaces" or "personas" that users can create and manage. The default state of a new conversation could be context-free of "incognito", requiring explicit user action to engage a persistent memory workspace. The system also could adopt explicit controls for activation, providing clear mechanisms to select which workspace is active for a given conversation, allowing the user to authoritatively add and use memory.

The interface layer: scaffolding understanding through transparency

The system should provide an interactive feature that functions as a transparent communication interface of its memory. Each memory entry should be easily auditable, with its origin clearly noted (e.g., "summarized from our conversation"). This demystifies how memory is constructed and combats user skepticism about hidden processes. When a memory influences a response, it could be surfaced directly within the conversational interface. This can be achieved through non-intrusive UI elements like footnotes or tooltips that explicitly state why a piece of information is being used.

The interaction layer: enabling co-curation through negotiated agency

On a foundation of sound architecture and a transparent interface, the interaction could be redesigned as a collaborative dialogue. Systems should not unilaterally decide what to remember, a process users found sometimes error-prone. For memories the systems generates, it could enter a "pending" state, prompting the user with a "review and commit" workflow to approve, edit, or reject the proposed memory before storage. When new information contradicts a stored memory, the system could also flag the discrepancy and ask the user for guidance. Similarly, when the system makes a significant inference, it could also seek confirmation, treating its own conclusions as hypotheses to be validated by the user, not as facts.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant No. 62472243 and 62132010.

References

[1] Sumit Asthana, Jane Im, Zhe Chen, and Nikola Banovic. 2024. " I know even if you don't tell me": Understanding Users' Privacy Preferences Regarding AI-based Inferences of Sensitive Information for Personalization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[2] Shubhi Asthana, Bing Zhang, Ruchi Mahindru, Chad DeLuca, Anna Lisa Gentile, and Sandeep Gopisetty. 2025. Deploying Privacy Guardrails for LLMs: A Comparative Analysis of Real-World Applications. *arXiv preprint arXiv:2501.12456* (2025).

[3] Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep Me Updated! Memory Management in Long-term Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 3769–3787.

[4] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. 2012. The menlo report. *IEEE Security & Privacy* 10, 2 (2012), 71–75.

[5] Natã M Barbosa, Zhuohao Zhang, and Yang Wang. 2020. Do Privacy and Security Matter to Everyone? Quantifying and Clustering {User-Centric} Considerations About Smart Home Device Adoption. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. 417–435.

[6] Tom L Beauchamp et al. 2008. The belmont report. *The Oxford textbook of clinical research ethics* (2008), 149–155.

[7] Ann Cavoukian et al. 2009. Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada* 5, 2009 (2009), 12.

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

[9] Giana Eckhardt. 2002. Culture's consequences: Comparing values, behaviors, institutions and organisations across nations. *Australian journal of management* 27, 1 (2002), 89–94.

[10] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6491–6501.

[11] Google PAIR. 2019. *People + AI Guidebook*. Technical Report. Google Research. https://design.google/aiguidebook/

[12] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen Macneil. 2023. Memory Sandbox: Transparent and Interactive Memory Management for Conversational Agents. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 97, 3 pages. doi:10.1145/3586182.3615796

[13] Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. 2020. Privacy concerns in chatbot interactions. In *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers 3*. Springer, 34–48.

[14] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models.

Advances in Neural Information Processing Systems 36 (2023), 20750–20762.
[15] Robert S Laufer and Maxine Wolfe. 1977. Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of social Issues* 33, 3 (1977), 22–42.
[16] Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. 2024. PrivLM-Bench: A Multi-level Privacy Evaluation Benchmark for Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 54–73.
[17] Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, et al. 2024. LLM-PBE: Assessing Data Privacy in Large Language Models. *Proceedings of the VLDB Endowment* 17, 11 (2024), 3201–3214.
[18] Yao Li, Alfred Kobsa, Bart P Knijnenburg, and MH Carolyn Nguyen. 2017. Cross-cultural privacy prediction. *Proceedings on Privacy Enhancing Technologies* (2017).
[19] Hannah Limerick, David Coyle, and James W Moore. 2014. The experience of agency in human-computer interactions: a review. *Frontiers in human neuroscience* 8 (2014), 643.
[20] Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. 2023. Think-in-Memory: Recalling and Post-thinking Enable LLMs with Long-Term Memory. *CoRR* abs/2311.08719 (2023). http://dblp.uni-trier.de/db/journals/corr/corr2311.html#abs-2311-08719
[21] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.
[22] Rongjun Ma, Caterina Maidhof, Juan Carlos Carrillo, Janne Lindqvist, and Jose Such. 2025. Privacy Perceptions of Custom GPTs by Users and Creators. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.
[23] Ying Ma, Shiquan Zhang, Dongju Yang, Zhanna Sarsenbayeva, Jarrod Knibbe, and Jorge Goncalves. 2025. Raising Awareness of Location Information Vulnerabilities in Social Media Photos using LLMs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–14.
[24] George R Milne, George Pettinico, Fatima M Hajjat, and Ereni Markos. 2017. Information sensitivity typology: Mapping the degree and type of risk consumers perceive in personal data sharing. *Journal of Consumer Affairs* 51, 1 (2017), 133–161.
[25] Helen Nissenbaum. 2011. A contextual approach to privacy online. *Daedalus* 140, 4 (2011), 32–48.
[26] Judith S Olson, Jonathan Grudin, and Eric Horvitz. 2005. A study of preferences for sharing and privacy. In *CHI'05 extended abstracts on Human factors in computing systems*. 1985–1988.
[27] OpenAI. 2024. Memory and New Controls for ChatGPT. https://openai.com/blog/memory-and-new-controls-for-chatgpt. Accessed: 2024-08-30.
[28] Arielle Pardes. 2018. The Emotional Chatbots Are Here to Probe Our Feelings. *Wired* (oct 2018). https://www.wired.com/story/replika-open-source/
[29] Hanna Schneider, Florian Lachner, Malin Eiband, Ceenu George, Purvish Shah, Chinmay Parab, Anjali Kukreja, Heinrich Hussmann, and Andreas Butz. 2018. Privacy and personalization: the story of a cross-cultural field study. *Interactions* 25, 3 (2018), 52–55.
[30] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. [n. d.]. PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
[31] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*. PMLR, 31210–31227.
[32] Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. 2023. Recursively summarizing enables long-term dialogue memory in large language models. *arXiv preprint arXiv:2308.15022* (2023).
[33] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 214–229.
[34] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136.
[35] Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Quanquan Gu, et al. 2024. Large Language Models Can Be Contextual Privacy Protection Learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 14179–14201.
[36] Anran Xu, Zhongyi Zhou, Kakeru Miyazaki, Ryo Yoshikawa, Simo Hosio, and Koji Yatani. 2024. DIPA2: An Image Dataset with Cross-cultural Privacy Perception Annotations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–30.
[37] J Xu. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567* (2021).
[38] Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long Time No See! Open-Domain Conversation with Long-Term Persona Memory.
[39] Yaqing Yang, Tony W Li, and Haojian Jin. 2024. On the Feasibility of Predicting Users' Privacy Concerns using Contextual Labels and Personal Preferences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.
[40] Ryan Yen and Jian Zhao. 2024. Memolet: Reifying the Reuse of User-AI Conversational Memories. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–22.
[41] Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. 2024. FinMem: A performance-enhanced LLM trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, Vol. 3. 595–597.
[42] Bo Zhang and S Shyam Sundar. 2019. Proactive vs. reactive personalization: Can customization of privacy enhance user experience? *International journal of human-computer studies* 128 (2019), 86–99.
[43] Shuning Zhang, Ying Ma, YongquanOwen' Hu, Ting Dang, Hong Jia, Xin Yi, and Hewu Li. 2025. From Patient Burdens to User Agency: Designing for Real-Time Protection Support in Online Health Consultations. *arXiv preprint arXiv:2508.00328* (2025).
[44] Shuning Zhang, Lyumanshan Ye, Xin Yi, Jingyu Tang, Bo Shui, Haobin Xing, Pengfei Liu, and Hewu Li. 2024. " Ghost of the past": identifying and resolving privacy leakage from LLM's memory through proactive user interaction. *arXiv preprint arXiv:2410.14931* (2024).
[45] Shuning Zhang, Xin Yi, Haobin Xing, Lyumanshan Ye, Yongquan Hu, and Hewu Li. 2024. Adanonymizer: Interactively Navigating and Balancing the Duality of Privacy and Output Performance in Human-LLM Interaction. *arXiv preprint arXiv:2410.15044* (2024).
[46] Zhiping Zhang, Bingcan Guo, and Tianshi Li. 2024. Privacy Leakage Overshadowed by Views of AI: A Study on Human Oversight of Privacy in Language Model Agent. *arXiv preprint arXiv:2411.01344* (2024).
[47] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.
[48] Xiangyu Zhao, Longbiao Wang, and Jianwu Dang. 2022. Improving dialogue generation via proactively querying grounded knowledge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6577–6581.
[49] Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. *arXiv preprint arXiv:2204.08128* (2022).
[50] Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. 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*. 1–28.
[51] Noé Zufferey, Sarah Abdelwahab Gaballah, Karola Marky, and Verena Zimmermann. 2025. "AI is from the devil." Behaviors and Concerns Toward Personal Data Sharing with LLM-based Conversational Agents. *Proceedings on Privacy Enhancing Technologies* 2025, 3 (2025), 5–28.

A Ethical Considerations

We followed Menlo report [4] and Belmont report [6] in considering the ethical implications. Notably, our study got the approval of our institution's Institutional Review Board (IRB). Participants are informed of the aim of the experiment, asked to sign the consent form before participating the experiment, and informed that they could withdraw the experiment at any time without reasons. Our study's aim is to facilitate more privacy-aware usage of LLM's memory through understanding users' mental models, practices and challenges.

B Interview Script

The original questions are in Chinese. We translate them to English without altering their meanings. During the study, we encouraged the users to reflect on their current memories, and their chat and memory histories.

B.1 Perception of RAG-based LLM Memory

We first provided a short description of Retrieval Augmented Generation, to prevent the case that participants did not understand this term, and we ensured participants' understanding before proceeding.

- In your view, how does the memory mechanisms of the a Large Language Model (e.g., ChatGPT) operate?
- Whether or not you think large model can sometimes remember information from your previous conversations? (And if yes, could you provide a specific example of when this has happened?)
- What information do you think it would remember?
- Whether or not you think it will remember private information? (And if yes, have you experienced some?)
- What do you see as the potential benefits of an LLM's memory function?
- Have you personally experienced any of these benefits? If so, could you describe the situation?
- What potential privacy risks do you associate with an LLM's memory function?
- Have you personally encountered a situation that you perceived as privacy risk? If so, could you describe it?
- How do you weigh the benefits and privacy risks?

B.2 Usage, Practice and Challenges

- How would you currently use the memory function of LLMs? Could you explain your reasoning?
- [Regarding memory generation] What is your current perception on how a memory is created? And what is your current behavior during this process?
- [Regarding memory generation] Is there any challenges during the memory generation process? (If so, please describe cases.)
- [Regarding memory management] What is your current perception on the memory management process? And what is your current behavior?
- [Regarding memory management] Is there any challenges during the memory management process? (If so, please describe cases.)
- [Regarding memory usage] What is your current perception on the memory usage process? And what is your current behavior?
- [Regarding memory usage] Is there any challenges during the memory usage process? (If so, please describe cases.)
- [Regarding memory update] Whether or not you have noticed the update of memory? And if so, what is your current behavior?
- [Regarding memory update] Is there any challenges during the memory update process? (If so, please describe cases.)

B.3 Perceptions of Inference in Memories

- Whether or not you believe that AI models can infer personal information from your past inputs?
- Whether or not there are any privacy concerns of the inference for this personal information? (If so, please describe.)
- If you think AI could infer things about you, what's your current behavior, and whether or not there are any mitigation. (If so, please describe)
- Are there any challenge of your mitigation strategies? And what's your expectation?

C User Consent

We showed a paper version of the user consent before the study. The original consent is in Chinese and we translated it to English without altering its meaning.

We are a research group from XX institution, investigating on users' perception of RAG-based LLM memory. RAG-based LLM memory is a form of memory that memorizes users' past preferences, personal interests or other personal information, that could be used in the future for enhancing conversation quality. It is evident in ChatGPT and other products. Our study's focus is to understand your perception on the RAG-based LLM's memory, your current practices and challenges, as well as your viewpoints on the potential inference behavior.

The interview would take approximately 30-60 minutes depending on its content, and would be audio recorded, and transcribed for academic analysis and publication. We would not use your material including the audio and transcribed text for any other usage than outlined above. Your participation is completely voluntary, and you has the right to withdraw at any time without penalty or explanation. Your data would be kept confidential and anonymized before processing. If you complete the experiment, you could get compensation according to the local wage standard (100RMB).

If you have any other questions, you could contact XXXX (Email: XXXX) for further clarification.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009