"Ghost of the past": Identifying and Resolving Privacy Leakage of LLM's Memory Through Proactive User Interaction

SHUNING ZHANG*, Institute for Network Sciences and Cyberspace, Tsinghua University, China

LVMANSHAN YE*, Shanghai Jiao Tong University, China

XIN YI†, Institute for Network Sciences and Cyberspace, Tsinghua University, China and Zhongguancun Laboratory, China

JINGYU TANG, Huazhong University of Science and Technology, China

BO SHUI, Shenzhen International Graduate School, Tsinghua University, China

HAOBIN XING, Tsinghua University, China

PENGFEI LIU, School of Electronic, Qingyuan, Shanghai Jiao Tong University, China

HEWU LI, Tsinghua University, China

Memories, encompassing past inputs in context window and retrieval-augmented generation(RAG), frequently surface during human-LLM interactions, yet users are often unaware of their presence and the associated privacy risks. To address this, we propose MemoAnalyzer, a system for identifying, visualizing, and managing private information within memories. A semi-structured interview(N=40) revealed that low privacy awareness was the primary challenge, while proactive privacy control emerged as the most common user need. MemoAnalyzer uses a prompt-based method to infer and identify sensitive information from aggregated past inputs, allowing users to easily modify sensitive content. Background color temperature and transparency are mapped to inference confidence and sensitivity, streamlining privacy adjustments. A 5-day evaluation(N=36) comparing MemoAnalyzer with the default GPT setting and a manual modification baseline showed MemoAnalyzer significantly improved privacy awareness and protection without compromising interaction speed. Our study contributes to the growing field of privacy-conscious LLM design, offering insights into user-centric privacy protection for Human-AI interactions.

Additional Key Words and Phrases: Large Language Models, Memory, Privacy Inference, Privacy Awareness

1 INTRODUCTION

The widespread use of Conversational Agents (CAs) based on Large Language Models (LLMs) has facilitated natural language communication while simultaneously posing significant challenges to users' privacy [24]. To enhance user experience and minimize repeated input of task-related background information, LLM service providers, such

---

*Both authors contributed equally to this work.

†This is the corresponding author.

as OpenAI, have implemented various memorization techniques. These methods involve deducing and storing user information based on natural language interactions[82, 97]. The powerful inference capabilities of LLMs, combined with increasing user data disclosure [74, 93], embed significant private information in memory traces that persist indefinitely unless explicitly deleted by the userthemselves [1]. As interactions with LLM-powered chatbots become more frequent across all aspects of life, the threat of privacy-invasive chatbots has risen, particularly concerning memory. Moreover, memory generation and usage in LLMs often lack transparency [97], frequently occurring without user consent or awareness.

Long-term memory operates using a retrieval-augmented generation (RAG-based) method, while short-term memory retains past user input in the context window. These approaches mimic human memory mechanisms [69] and together form the LLM's "memory" of user input. Both memorized content could be leveraged for training, posing significant privacy leakage risks [75]. This opacity leads to inadvertent user consecutive contributions to memory systems, heightening privacy risks [63] and vulnerability to membership inference attacks [75]. For example, You have told ChatGPT about your preferences for work hours and life balance. Recently, you have asked ChatGPT for help about the difficulties you encounter at work. When brainstorming the possibility of changing jobs or starting a business. ChatGPT may infer dissatisfaction with your job and plans for a career change based on prior conversations.

This research builds on previous classifications [92] to analyze memory-related risks and corresponding countermeasures across memory generation and usage phases. Our focus is on privacy leakage risks [40] and user inference attacks [35], both critical threats in LLMs [35, 51]. Two main privacy risks arise from user inputs when transferred to memory: (1) individual inputs or long-term memories may contain sensitive information, and (2) the aggregation of these past inputs and memories may lead to the exposure of sensitive data. When utilized for model training or fine-tuning, these inputs pose a significant privacy threat by potentially enabling the model to infer and expose personal information [12, 16].

The past input and memory consisting sensitive information, as well as the inferred private information both presented risks of privacy leakage [40, 51]. The persistent use of such memory intensifies these risks, yet privacy implications remain under-explored, with limited research on risk categorization and mitigation. Hence, we aim to answer this research question: **How to design a transparent and controllable notification technique which timely triggers participants' awareness and mitigates potential risks in the memories of LLMs-based CA?**

To explore these questions, we first conducted a semi-structured formative interview study (N=40) to examine users' privacy perceptions towards LLM memory systems and their expectations for efficient memory management. The results showed that most users were unaware of the existence of memory systems, particularly long-term RAG-based memory, with only 5 out of 40 participants demonstrating an understanding of long-term mechanisms. Even users with frequent usage held misconceptions, such as believing memory was limited to a specific dialogue or could be shared with other users. After our explanation, participants expressed the need for transparent, controllable designs for visualizing and modifying the memory mechanism. We found that current LLM products already retain some personal privacy information in their memory, which accumulates over time. However, the inference process for this private information remains opaque, and users often only realize privacy risks after a delay, when tracing the original input becomes significantly more difficult.

To address the challenges identified in the formative study, we developed MemoAnalyzer, a pop-up browser plug-in that visualizes inferred private information and enables users to modify it easily. MemoAnalyzer appears after each interaction, distinguishing private information based on inference confidence and sensitivity [50, 52]. Inference confidence is represented through varying opacity levels, while sensitivity is indicated by color, with red signaling highly sensitive information and blue indicating lower sensitivity. When users click on a piece of

---

[1]https://openai.com/index/memory-and-new-controls-for-chatgpt/, accessed by Sep 12th, 2024

private information, MemoAnalyzer displays the relevant past inputs and memories, highlighting the keywords that contributed to the inference. This helps users quickly identify and modify the terms that contributed to the inference. By utilizing prompt-based inferences, MemoAnalyzer provides users with proactive control over their inputs and memory data. Since it only performs inference without training on the data until users take action, privacy risks are minimized at this stage. Once users modify or delete their data, any potential risk from future model training is effectively mitigated.

A five-day in-lab evaluation study (N=36) demonstrated the effectiveness of MemoAnalyzer compared to *GPT* implementation and *Manual* management in three typical types of tasks: work-related, life-related and academic-related [93]. MemoAnalyzer was preferred for its comparable time efficiency, superior privacy protection, and enhanced user experience regarding perceived control, transparency, trust, and overall preference. Users particularly appreciated its flexibility, control-ability, and transparency. We envision MemoAnalyzer to be a superior solution for managing personal memory in LLMs, especially concerning sensitive information.

In summary, this work made three key contributions:

- We unveiled users generally lacked the timely and clearly awareness of the long-term memory mechanism in contrast to the short-term context memory through a interview study (N=40).
- We proposed MemoAnalyzer, a technique notifying users' privacy risk, enabling users to selectively control their private information. MemoAnalyzer facilitated the collaborative privacy information management where users indicate their preference.
- We evaluated MemoAnalyzer in an user study (N=36) compared with GPT and Manual settings, where MemoAnalyzer were favored for its comparable speed, higher privacy protection capability and higher user experience. It also enabled users to control their privacy more transparently.

2 RELATED WORKS

We first introduced the privacy risk in LLM-based CAs. Then we detailed the memory mechanisms of LLMs and the potential privacy risk. Finally, we categorized the privacy awareness of end-users in human-AI interaction.

2.1 Privacy Risk in LLM-based CAs

With a service targeting conversational assistants, we detail privacy challenges in LLM-based CAs memorization. To optimize conversational performance, LLMs inherently require vast amounts of data for their training, often encompassing user interaction data [62]. However, a side effect of LLMs is the unintentional memorization of the training data, which also contain user input data, including personally identifiable information (PII) [12, 66], which might also be included in the generated output. For example, ChatGPT, even with safety precautions, can inadvertently disclose PII through specifically crafted prompts.

Users engage with LLM-based CAs through natural language, which is traditionally reserved for human-to-human communication. This can lead them to perceive these agents as human-like. Studies suggested that anthropomorphizing can increase user information disclosure [34, 41]. Anthropomorphizing can inflate users' perceptions of the CAs' competencies, fostering undue confidence, trust, or expectations in these agents [41, 100]. With more trust, users might be more inclined to share private information, even in contexts typically associated with sensitive personal information [41, 80, 100]. Anthropomorphization may amplify the risks of users yielding effective control by trusting CAs unquestioningly. Moreover, more private information may be revealed when CAs leverage psychological effects, such as nudging or framing [84].

Generalized to AI technologies, the past literature [42] classified the risks as invasion risks [55, 61, 67, 73], data collection risks [73], data dissemination risks [4, 13, 44] and data processing risks [64, 65, 72, 73, 77]. Invasion risks encompass a range of activities that intrude upon an individual's personal space or solitude. Intrusion risks encompass actions that disturb one's solitude in physical space [73], which include personalized ads. Besides,

surveillance is common with the support of AI technologies [61, 67], and the ubiquity of sensors [55]. Data collection risks "create disruption based on the data gathering process" [73], which exacerbated surveillance risks [73], further exacerbating surveillance risks. Data processing risks result from the use, storage and manipulation of personal data [72, 73].

2.2 Memory in LLMs

With the increasing complexity of human-AI interaction [87] and the tasks [2], memory becomes significant during the interaction [5, 31]. Current LLMs remain opaque about the usage of memory [31] and the context [85, 94]. To address these challenges, researchers have explored various strategies to manage memory more effectively and transparently. Some proposed techniques for retaining the persona of chatbots [45, 91]. Other methods guarantee the responses generated are contextually appropriate, such as summarizing [81] and refinement [95], aiming to minimize redundancy while maintaining essential information. Relevant memories can be retrieved utilizing information retrieval techniques to contextualize current inputs to AI [5, 86]. One of the critical aspects of privacy risks in LLM-based systems is the memory mechanism these models employ to retain and utilize user data across interactions. Memory in LLMs can be broadly categorized into short-term memory, which involves retaining user input for the duration of a session, and long-term memory, which stores user interactions across multiple sessions to enhance continuity and personalization [98]. Recent studies have highlighted the opacity and complexity of memory mechanisms in LLMs, which often operate without explicit user consent or understanding [70]. For example, research showed that users are typically unaware of how LLMs store and use their data, leading to significant privacy concerns, especially regarding the potential for long-term retention of sensitive information [47].

This lack of transparency raises ethical concerns and poses substantial risks of data breaches and unauthorized access to private information [76]. Researchers have proposed various strategies for managing memory in LLMs to address these challenges, focusing on enhancing transparency and user control. One approach involves using external memory management systems that allow users to access, modify, or delete stored data proactively [53]. These systems often employ visualization techniques to help users understand what data has been retained and how it might be used in future interactions [32]. For instance, Huang et al. developed a memory sandbox tool that provides a transparent interface for managing conversational histories, enabling users to selectively edit or remove entries that contain sensitive information [39]. Moreover, advancements in prompt-based memory management have been explored to facilitate better control over the data retained by LLMs. This involves using structured prompts to guide the model in identifying and managing relevant information while minimizing the retention of unnecessary or sensitive data [96]. Recent implementations of these techniques have demonstrated their effectiveness in reducing privacy risks and enhancing user trust by providing more granular control over memory retention and usage [46].

Despite these efforts, the process of "remembering" remains complex or machines [48, 81], which is even harder for humans to take control. Users often lack a clear understanding of how generative AI and conversational agents handle memories [31]. Current tools primarily focus on accessing and editing chat histories to manage conversational memories [31, 59].

2.3 User-centered Privacy in Human-LLM Interaction

Users' perception of privacy critically influences their disclosure behaviors and the potential for privacy leakage in human-LLM interactions. However, due to LLMs' opaque privacy management mechanisms, users often lack sufficient privacy awareness, undermining their ability to make informed decisions about personal data disclosure.

To address this pervasive issue, researchers have proposed various countermeasures within the "notice and control" paradigm. Yet, these measures frequently encounter significant limitations. Non-salient privacy notices fail to effectively capture users' attention, especially when concealed behind hyperlinks or embedded in click-wrapped agreements. For instance, Cate [14] highlighted that on Yahoo's website in 2002, a mere 0.3% of users read the click-wrapped privacy policy, a figure that increased to only 1% after a public privacy controversy [28]. Similarly, in an experiment with a fictitious search engine, none of the 120 participants accessed the privacy policy link [27]. In another study, only 26% of users joining a simulated social network viewed the policies [27], and in a survey scenario, just 20.3% of participants clicked to view privacy information [79].

Enhancing user control and transparency is therefore essential for fostering trust and safeguarding privacy in AI systems, including LLMs. Privacy-preserving techniques that empower users to modify or delete their personal information can substantially mitigate privacy risks [1]. Transparent AI systems that clearly communicate how data is collected, processed, and stored further strengthen user trust and promote responsible AI practices [21]. Adopting a user-centric approach to privacy design—emphasizing effective notice and granular control over personal data—has gained significant attention. Tools such as interactive privacy dashboards and real-time data usage notifications enable users to make informed decisions about their privacy [22]. Research indicates that users are more inclined to trust and engage with AI systems that provide clear explanations of data handling and allow for easy adjustments to privacy settings [71]. However, these systems lacked the analysis into the memory mechanism. Thus, this paper initiated the first study to visualize the privacy and enable users to modify the private information in memory.

3 STUDY 1: EXAMINING USERS' AWARENESS AND PRACTICE USING LLM'S MEMORY

In this section, we conducted an interview study to explore users' practices and privacy concerns regarding memory during their interactions with LLMs.

3.1 Study Design

The experiment was conducted through semi-structured interviews. The session began with questions about participants' demographics. Participants were then asked about their knowledge and experiences with the memory systems in LLM products. We then divided the memory usage into three stages according to the past literature [30] and OpenAI's official introduction[2]: memory generation, memory management and memory usage. For each stage, we explained the relevant implementation and usage mechanisms, then asked participants to discuss the advantages and dis-advantages of these mechanisms. After gathering their subjective opinions, we inquired about their expectations for each memory stage. The full interview script is provided in Section A of the appendix. Experimenters also posed additional open-ended questions if any unexpected insights emerged during the interviews.

3.2 Recruitment and Participants

This IRB-approved study recruited 40 Chinese participants (13 males, 27 females) with a mean age of 22.6 (SD=2.1) in XX campus (anonymized for submission) through snowball sampling [26]. To ensure the diversity of participants, we distributed the questionnaire in different chat groups in different hours of a day. We continued the recruiting while limiting the participants each day until the results saturated following saturation theory [23]. After saturation, we continued to recruit another 3 participants. Five participants were with high school education degree, 16 were with bachelor education degree, 17 were with master degree and 2 were with Ph.D. degree. Participants self-rated their familiarity towards the LLMs and AI, as well as their usage frequency, the result of which was shown in Figure 1. 7 participants were from design background and 6 were from computer science

---

[2]https://openai.com/index/memory-and-new-controls-for-chatgpt/

background, with no one from security and privacy background and other from other backgrounds. 39 participants have used ChatGPT [3], 19 participant have used Kimi Chat [4] and 12 participants have used WenXinYiYan [5]. All participants reported having used LLMs at least once. Each participant who completed the experiment received 90 RMB as compensation.
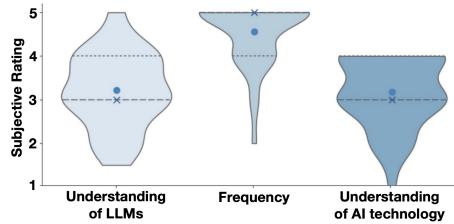

Fig. 1. Participants' familiarity and usage frequency towards AI (5: most familiar and frequent, 1: least familiar and frequent). The cross sign indicated the median and the square sign indicated the mean.

3.3 Procedure

The experiment was conducted via Online Meeting[6]. Participants were first briefed on the study and asked to provide informed consent. They were free to withdraw at any point during the experiment process. The entire experiment process lasted about 50 minutes on average and all recorded sessions were transcribed for analysis.

3.4 Analysis Methods

We used thematic analysis [11] to qualitatively code the results. As the study followed grounded theory, the codebook was iteratively refined throughout the process. The first two authors jointly coded all data, engaging in periodic discussions to resolve disagreements. The coding process incorporated open coding [37], axial coding [36] and inductive coding [15]. Initially, open coding was used to identify the primary set of codes, which were then grouped into axial codes and broader themes. Because of the inductive and iterative nature of the experiment, guided by the past literature [54], agreement score is not suitable for this scenario. We intended not to report the agreement score. We calculated the frequency of utterance after qualitative coding.

3.5 Results

We first revealed the cognitive gap of participants, then presented users' expectations towards future improvement. We detailed three findings shortened as F1 to F3.

3.5.1 F1: Memory Mechanism is Opaque. The process of memory generation within the system is problematic due to its opacity. Without explicit notification, 30/40 of users were un-aware of the notification when memory and context were added. The memory was also showed in the management interface in a plain manner, without highlighting the potential private information, resulting in no participant proactively noticing the privacy inside memories. The past input were even not showed in a structured panel, leaving no participant aware of the potential privacy leakage. Although all participants re-vealed private information could be directly contained in the past input or memory during generation, 37/40 participants also envisioned private information could be

---

[3]https://chatgpt.com/

[4]https://kimi.moonshot.cn/

[5]https://yiyan.baidu.com/

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

inferred from multiple past inputs or memories. Worse still, 28/40 users were totally un-aware of the memory usage. Even these 28 participants who were aware of the memory usage, they could not understand or guess out how the memories are leveraged and integrated. Participants also has no manner to effectively regulate or modify the memory. The management panel is hard to find, and 35/40 participants never clicked it. The 5 participants who clicked it exhibited simple behaviors. Three participants directly deleted all the memories because of the memory risk, whereas two participants chose to retain all the memories regardless of the privacy risk. Participants typically commented, *"I never considered GPT has this function before this day. I am definitely not known about the private information inferred from the memory."*

3.5.2 F2: Users Lack the Privacy Awareness towards the Memory Mechanism. Despite the fact that users were un-aware of the long-term memory generation, they were also un-aware of the private information contained or potentially inferred from the past input and the memory unless explicitly told. 30/40 participants would never notice the private information in the past input before the experiment, and unfortunately, 36/40 participants never notice the private information in the memories before the experiment. They were even more un-aware of the risk that the private information could be inferred through combining different past user input and memory, which was exactly done in the usage of the long-term memory. In fact, among the interviews, all participants echoed *"I have never thought about these information usage and leakage patterns before."* All participants were never aware of the private information inferred combining the past inputs and memories. However, interestingly, participants could understand the inference process after explicitly told. This further un-veiled the in-transparency of the system's memory management.

3.5.3 F3: Users Need the Control of the Memory Mechanism. All participants reflected they lacked of control over the memory mechanism, highlighting the necessity for proactive memory management. All participants expressed the need for greater autonomy in controlling their memory, which can be categorized into three key actions: editing, adding, or deleting memories. Notably, a considerable percentage of participants identified privacy as the primary motivator for managing their memories, with privacy concerns being the most critical factor (35/40), followed by the accuracy of the stored information (14/40). 36/40 participants commented *"I definitely need the system to provide me with the proactive modification permission."* A few participants also hoped besides proactive modification, the system could automatically help them manage private information. When it comes to private information, users indicated the need to make decisions about deleting such data based on the usage, importance, and potential privacy risks associated with the memory. Specifically, categories such as health information, academic information, and personal basic information were frequently deleted due to their sensitivity, whereas categories like preference, formatting options and research interests were often retained to enhance usability.

4 DESIGN AND IMPLEMENTATION OF MEMOANALYZER

MemoAnalyzer is designed to address privacy concerns aside users' tasks. It analyzes user-added memories and inputted information, providing a visualization for users to review, modify, or delete the data. This process is done before the real fusion of different memory and inputted information source during hypothetical later input (usually during training). Deletions are performed post hoc, striking a balance between preserving model output performance and safeguarding privacy, as the user's task has already been completed at that point.

4.1 Design Goals

We proposed several design goals according to the formative study:
DG1: Enhance the transparency of (private) memory information inferred from user input and past memory through visualization. (F1) Users were seldom aware of the private information inference, especially

combining past input and memory together. Visualizing is an effective method to provide information to users intuitively [19].

DG2: Increase users' awareness of (private) memory inference through highlighting the inferred text and the original keywords. (F2) Few users were aware of the memory inference process, especially the keywords involved in the inference. Additionally, the private information is not transparently illustrated. Thus, the visualization of the keywords could make users aware of the opaque process [83].

DG3: Support users' proactive control of (private) memory information such as modifying and selectively adding. (F3) Often, the system could automatically conduct the modification. However, towards the privacy content, users needed to have the proactive control, which users could selectively protect their privacy according to their preference and further propelled the forming of privacy awareness [25].

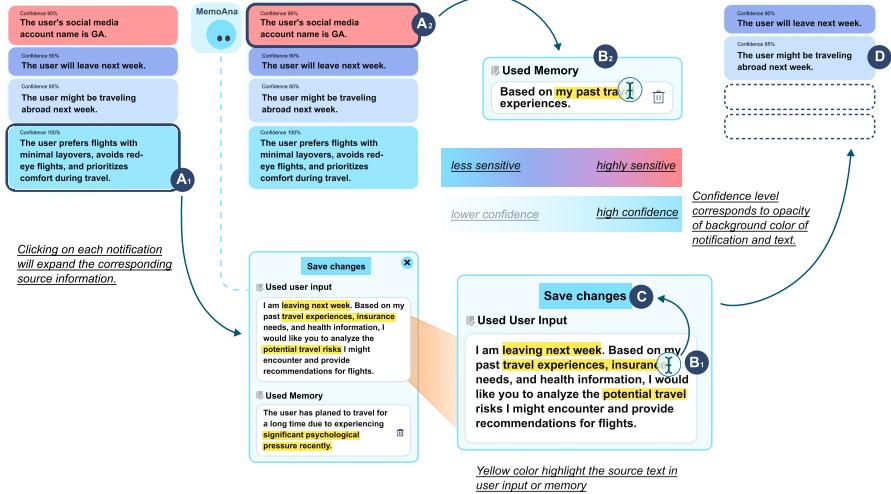## 4.2 Interface and Interaction Flow

Fig. 2. MemoAnalyzer's different functions. (A) The user click the notification attracting their curiosity. (B) The past inputs and memories used to infer the private information are expanded below. The specific phrases used for inference are highlighted to facilitate users' modification. (B1, B2) Users can edit or delete the memories while edit the past input. (C) User clicks the "save changes" button after modification to save the changes. (D) The inferred private information disappears or changes after user's modification.

MemoAnalyzer adopted a minimal interface design [38] in the form of a pop-up, as shown in Figure 2. It followed the previous intervention design, placing the interface in the peripheral (the left) of the interface, reducing the disturbance to users' interaction. Upon completing the input, the users could view the private information inferred from the past inputs and memory histories, as demonstrated on the left of the interface (see Figure 2 (A)). The main elements in the intervention interface are the inferred private information based on the past input and the memory. We chose to demonstrate all inferred private information to increase users' control-ability and enhance their privacy awareness. MemoAnalyzer used different highlight according to the confidence of the inference and the sensitivity of the information. The confidence of the inference was defined as "how sure the LLMs think the inferred information is?". The confidence of the inference $c \in [0, 1]$ controls the

transparency of the block, while the sensitivity of the information $s \in [0, 1]$ controls the color of the block. The more sensitive the information, the redder the information was highlighted. The relative privacy was determined through a questionnaire-based rating pilot study with the methods similar to the past literature about online text privacy [9, 29]. We transformed the original mean ratings to between 0 and 1 using linear mapping. The color was determined by the following equation:

$$rgba(c, s) = (109 + s * (255 - 109), 172 + s * (117 - 172), 255 + s * (117 - 255), c) \tag{1}$$

We determined not to show the original text by default to reduce users' mental load of viewing the privacy information. Only when users proactively click the private information would the system show the information source to users. To reduce the cognitive load of users, we designed a hierarchical interaction approach. Notifications are given first, followed by detailed expansion. Clicking on the private information inferred would unfold the original text source and users could view the input and memory history (see Figure 2 (B)). The information source was divided into the past input (short-term memory) and the past memory (long-term memory). The past memory supported editing and deleting, while the past input supported editing. The keywords used to infer the private information were highlighted in yellow color (see Figure 2 (D)). We used the direct click to indicate the start of the editing, aligning with previous literature [99]. Users could edit on the history, delete or edit the memory and click "Save Changes" to record the modification (see Figure 2 (D)).

4.3   Implementation of MemoAnalyzer

We implemented MemoAnalyzer using Javascript and Python, where the main notification floating window is implemented as a plug-in. The frontend and backend adopted Flask framework. For the memory management and the inference, we used a one-shot prompting method, similar to GPT's memory generation and custom instruction processes [56, 57]. This approach requires no user data training, ensuring participants' privacy. The implementation is detailed in the following sections, *with additional information provided in the supplementary materials.*

4.3.1   Privacy Inference. To better protect users' privacy, we used the most advanced method of inference [76]. We inferred the potential private information based on all the past user input in the current dialogue and all the past memories. The privacy inference required the system to "infer and identify the personal sensitive information" from "the past inputs and the memory" as much as possible. We also added prompts facilitating sensitivity highlighting and source tracking, which we detailed in the following sub-sections. We also added the prompts to reduce repetitive private information. We prompted the LLMs (using GPT-4o-2024-05-13) to output in a structured manner and extracted all the private information items to demonstrate on the interface. The input consisted of the step-by-step description, the definition of different private information types, rules, formats (including the private information, its type, the confidence, the original past inputs, the original memory) and the one-shot example. The one-shot example was manually crafted by experimenters. The structured output is a list of multiple private information, along with their confidence, types and the original text.

4.3.2   Sensitivity Highlighting. We opted to output the confidence and sensitivity of the information alongside the inference. This choice is driven by 1) the latency would be reduced to a simple query, and 2) the LLMs were tested capable of handling the information and sensitivity together. The confidence was prompted to output with the private information. We also prompted the LLMs to output the sensitive information type. LLMs was asked to output the corresponding type along each private information. We used the sensitivity rating in the literature [9, 29] to represent the sensitivity of the specific private information, and visualized them through the transition from red to blue.

4.3.3 Source Tracking. We used a prompt-based method alongside the previous steps for tracking the source of the information. We let LLMs to output the source information used to infer the privacy and especially the keywords used for inferring the privacy. The LLMs was asked to separately output the past input and the memory, along with the keywords. We also tagged the unique identifier of each input and let LLMs select the unique identifier to facilitate later modification and replacement.

4.3.4 Editing Proxy. We proxied users' editing. Users edited on the left of the screen and the corresponding input as well as the memory would be modified correspondingly. This was achieved through assigning a unique id for each input, as mentioned in the last sub-section. Upon users submitting the modification, we tracked the modification and replaced the original history and the memory according to the unique identifiers through searching. This largely reduced users' time of viewing through all the past histories and modify manually.

5 STUDY 2: EVALUATING MEMOANALYZER

We conducted a five-day in-lab study to evaluate the effectiveness of MemoAnalyzer compared with other memory management methods.

5.1 Participants and Apparatus

We recruited 36 participants (12 males, 24 females) with a mean age of 23.9 (SD=2.7) from the XX campus (anonymized for submission) through snowball sampling [26]. Participants reported moderate familiarity with LLM products (M=4.33, SD=0.47) and usage frequency (M=3.66, SD=1.24), but lower familiarity with AI techniques (M=3.33, SD=0.47), indicating regular use of LLMs but limited understanding of underlying AI technologies. Participants were also not familiar with privacy and security researches and techniques. No participant dropped the experiment and each participants received $15 as compensation. The experiment was conducted through an online platform provided as a web service through Flask. Participants used their own laptop to connect to the provided website for the experiment to better mimick their own usage case. Experimenters connected to participants through online meeting[7].

5.2 Experiment Design

Since the memory patterns of LLMs do not fully emerge in a single day, we conducted a five-day study, focusing on short-term memory each day and analyzing long-term memory across the entire period. The period aligns with the previous literature [43, 90] and is proved to be efficient by the results see Section 5.4.

We used a one-factor within-subjects design with **technique** as the only factor. We compared MemoAnalyzer with two other techniques (see Figure 3):

- MemoAnalyzer: we implemented MemoAnalyzer according to the design and implementation section. Besides, we implemented the memory management interface as GPT products.
- GPT: we implemented the memory mechanism with a similar manner to GPT products[8] following official guidelines [60]. We used GPT to extract memory, to guarantee both extracting effect and a low latency. We implemented the same memory management interface as memoAnalyzer.
- Manual: we implemented the manual baseline with no context and memory, but a clipboard for users to manually copy and paste their past input. This method mimicking Temporary Chat [58]. They could also save the memory in the system similar to their local memorandum.

For all techniques, the back-end API was GPT-4o (version: GPT-4o-2024-05-13), the most advanced to facilitate comparison. We implemented the interface to support creating, managing dialogues for all systems. The chat

---

[7]https://meeting.tencent.com/
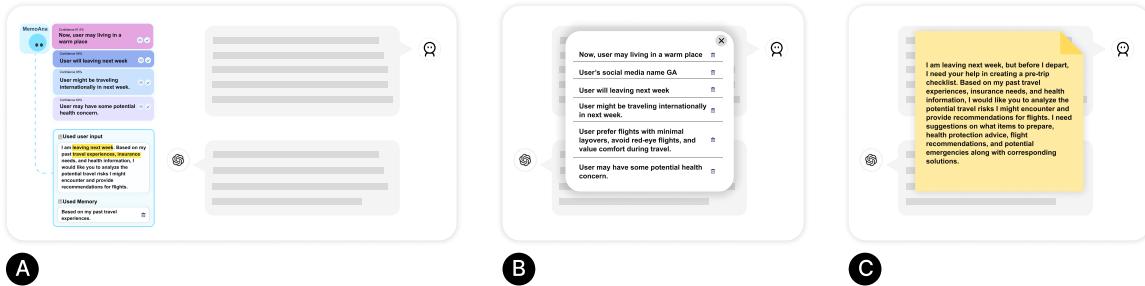[8]https://chat.openai.com

Fig. 3. The experiment platform of different techniques, (a) MemoAnalyzer, (b) GPT-4o, (c) Manual.

history would be maintained, however the context would never be used for the Manual baseline. For the memory mechanism of all baseline techniques and MemoAnalyzer, we followed the guidance of OpenAI for implementing. For the parts that OpenAI did not specify, we detailed in supplementary materials for its implementation following the past guidance in the literature [88]. We envisioned our memory mechanism is representative of the main stream LLMs and facilitate fair comparison. *The detailed content and prompt of the implementation were shown in the supplementary material.*

We selected three types of tasks related to users: work-related, life-related, and study-related [93]. Each task type was designed and distributed based on the ShareGPT90K dataset[9], commonly used for analyzing human-AI interaction. Tasks were carefully structured to be easily completed daily and to maintain correlations across days. Detailed task content is provided in supplementary materials. Participants were instructed to pseudo-anonymize their inputs to avoid sharing private information and mitigate potential ethical issues. Anonymization examples were provided for clarity.

Figure 4 showed the experiment design. Each participant completed three five-day tasks, varied by technique and task type, all within a single scenario. Scenarios were counterbalanced among participants. Tasks were divided to reflect daily usage patterns, ensuring each task correlated with previous ones and remained manageable for daily completion.

We used questionnaires and also analyzed users' behaviors, with measurements shown in Table 1. We also conducted 15-minute semi-structured exit interviews to all participants after the experiment of Day-1 and Day-5 (see Figure 4). The preset interview questions in the exit interviews included to let participants describe 1) how they use each functions in different systems to manage their memory and 2) how they perceive these functions regarding their usefulness and ease of use. The details of the exit interview was shown in the supplementary materials. Additionally, we asked participants about potential improvements of different techniques.

5.3 Procedure

We first informed participants of the experiment and gave them 3 minutes to become familiarize with the experiment platform. They then needed to sign the user consent before proceeding the study. We detailed the potential harm in the user consent and they were informed they could quit the experiment at any time. They needed to complete 3 sessions of experiment differed by the technique managing their memory and private information. Within each session, they needed to interact with the system for several turns until they were satisfied with their answer. They input the question in the input area of Figure 3 and clicked the submit button to see the private information processed in the same area in MemoAnalyzer. They followed the experiment

---

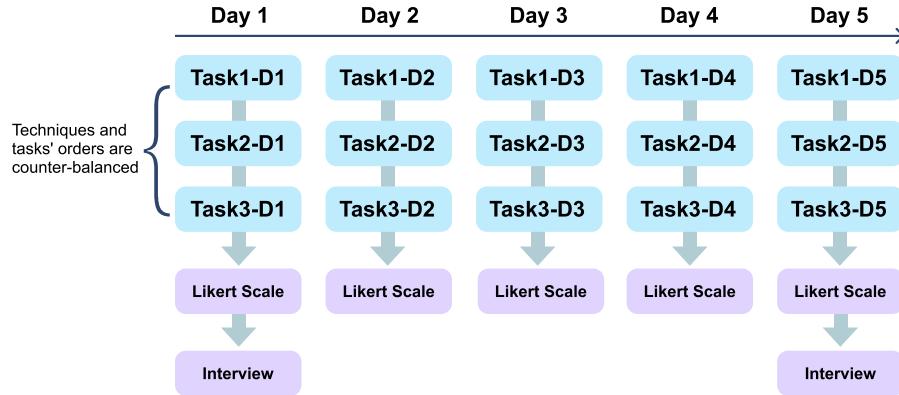[9]https://huggingface.co/datasets/liyucheng/ShareGPT90K

Fig. 4. An overview of the study's process. There are interviews on Day-1 and Day-5 separately and questionnaires after each day's tasks.

Table 1. Subjective and objective evaluation metrics for Study 2. The objective metrics were highlighted in blue.

Category	Metrics
Efficiency	**Total completion time** : time taken by users to complete tasks with each system.
	Privacy management time : time taken by users to manage their privacy. *There is no management time for Manual group.*
	Perceived privacy management speed: whether the system could fast and efficiently collaborate with you in managing the privacy. *There is no perceived privacy management speed for Manual group.*
Function usage	**Frequency** : measure how frequently participants used different functions such as selectively add, modify, delete the memories during memory generation and usage process.
User satisfaction	**Usability**: user satisfaction, ease of use and overall user experience.
	Control: whether users think they were in control of the private information.
	Transparency: whether the system transparent demonstrate the information to users.
	Effectiveness: whether the system could effectively handle privacy issues.
Cognitive load	**NASA-TLX**: NASA Task Load Index, cognitive load on users when interacting with the memory system.

guideline for other systems. They needed to complete their corresponding part of task for each technique for each day across five days. After the completion of each technique the participants needed to fill in the subjective evaluation questionnaire. All techniques were video-recorded. After Day-1 and Day-5 we asked participants to participate in a semi-structured interview, in which participants also needed to comment on how and why they

handle the private information. The experiment for each day lasted no more than 40 minutes and participants were each compensated 350RMB in total.

5.4 Results

We conducted statistical testing to all behavioral and subjective rating data. We performed Repeated Measures Analysis of Variance (RM-ANOVA) and Tukey post-hoc comparisons to behavioral data, whereas we performed Friedman non-parametric tests and Nemenyi post-hoc comparison to subjective rating data. We further performed thematic analysis [11] to subjective interviews. The themes were generated through a combination of open-coding [37] and axial-coding [36].
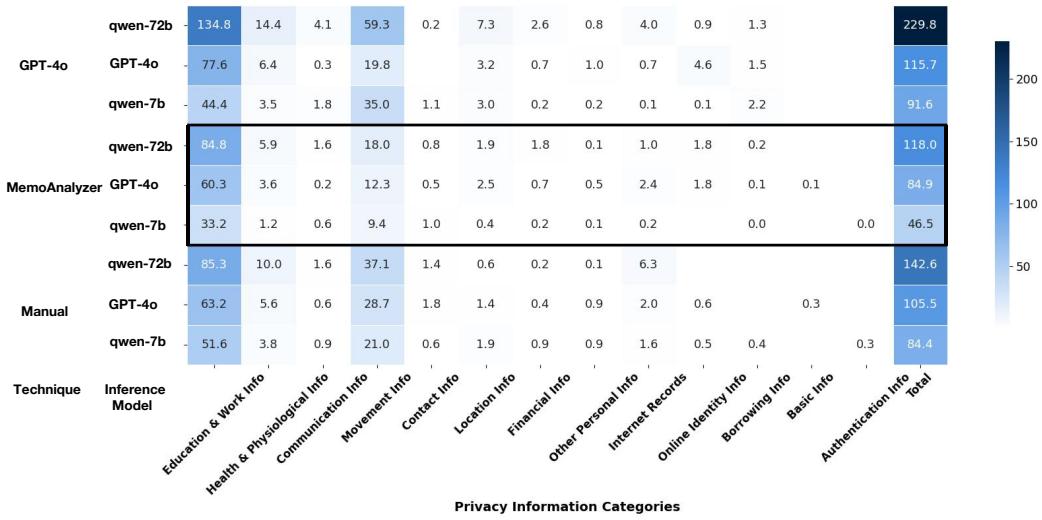
Technique	Inference Model	Education & Work Info	Health & Physiological Info	Communication Info	Movement Info	Contact Info	Location Info	Financial Info	Other Personal Info	Internet Records	Online Identity Info	Borrowing Info	Basic Info	Authentication Info	Total
GPT-4o	qwen-72b	134.8	14.4	4.1	59.3	0.2	7.3	2.6	0.8	4.0	0.9	1.3			229.8
	GPT-4o	77.6	6.4	0.3	19.8		3.2	0.7	1.0	0.7	4.6	1.5			115.7
	qwen-7b	44.4	3.5	1.8	35.0	1.1	3.0	0.2	0.2	0.1	0.1	2.2			91.6
MemoAnalyzer	qwen-72b	84.8	5.9	1.6	18.0	0.8	1.9	1.8	0.1	1.0	1.8	0.2			118.0
	GPT-4o	60.3	3.6	0.2	12.3	0.5	2.5	0.7	0.5	2.4	1.8	0.1	0.1		84.9
	qwen-7b	33.2	1.2	0.6	9.4	1.0	0.4	0.2	0.1	0.2		0.0		0.0	46.5
Manual	qwen-72b	85.3	10.0	1.6	37.1	1.4	0.6	0.2	0.1	6.3					142.6
	GPT-4o	63.2	5.6	0.6	28.7	1.8	1.4	0.4	0.9	2.0	0.6		0.3		105.5
	qwen-7b	51.6	3.8	0.9	21.0	0.6	1.9	0.9	0.9	1.6	0.5	0.4		0.3	84.4

Privacy Information Categories

Fig. 5. The heatmap of inferred information for different privacy information categories (the final column is the sum). The number denoted the private information item counts inferred using LLMs, averaged across participants. The horizontal axis denotes the technology (MemoAnalyzer, GPT-4o, Manual) and the LLMs used for inference (GPT-4o, qwen, qwen-7b). The numbers for MemoAnalyzer is outlined with black boundaries.

5.4.1 Privacy Protection Effectiveness. In response to the threat model where users' past inputs and memories are used for training and inference in LLMs, leading to privacy leakage, we tested three commercial LLMs—GPT-4o (estimated >100B), Qwen-72B, and Qwen-7B—to infer private information from users' utterances and memory. On the fifth day, we collected participants' past inputs and memory histories, conducting five inference attempts to minimize random effects. Due to the model API limit, we randomly chose 25 participants from total 36 participants (accounting over the half) to conduct the analysis. Each dialogue was analyzed along with its related memory, and results were aggregated across dialogues. We reported the average number of inferred private data per participant, categorized and reported separately [9, 29][10]. Figure 5 showed the number of different private information inferred using different LLMs. Specifically, for MemoAnalyzer, there was a marked reduction in the total amount of inferred private information. A statistical analysis revealed a significant difference in the total private inference for advanced models such as GPT-4o ($F_{2,48} = 4.35$, $p < .01$, post-hoc $p < .05$, compared with GPT and $p < .05$ compared with Manual) and qwen-72b ($F_{2,48} = 4.45$, $p < .01$, post-hoc $p < .05$ compared

---

[10]https://www.tc260.org.cn/upload/2021-12-31/1640948142376022576.pdf

with GPT and $p < .05$ compared with Manual). Even compared with models such as qwen-7b, there is still a significant difference ($F_{2,48} = 2.67$, $p < .05$), although post-hoc comparisons found no significant differences. This demonstrated the superior privacy protection capability of MemoAnalyzer.

Participants were found to frequently input education and work information, indicated by its highest frequency. MemoAnalyzer has its pronounced effect in guarding participants' privacy, with over 22.3% and 4.6% percentage of private information reduction evaluating using GPT-4o. Besides, participants also input less private information in these less frequent categories, resulting in a less total number.

5.4.2 Interaction Time. We defined total time as the duration from the first character input to the moment the website was closed, indicating task completion. Privacy protection time was measured by participants' interactions with notifications or memory panels for privacy management. Figure 6 presents the total and privacy protection times for each technique. Total time comprised both privacy protection and pure task completion time. *MemoAnalyzer* exhibited comparable total times (M=460.3s, SD=58.6s) to *GPT-4o* (M=426.2s, SD=50.6s) and *Manual* (M=462.7s, SD=55.6s). Similarly, the privacy protection time for *MemoAnalyzer* (M=29.9s, SD=6.6s) was close to *GPT-4o* (M=23.9s, SD=4.3s). No significant differences in total time were observed across techniques, except on Day 1, where both *MemoAnalyzer* and *GPT-4o* outperformed *Manual*.

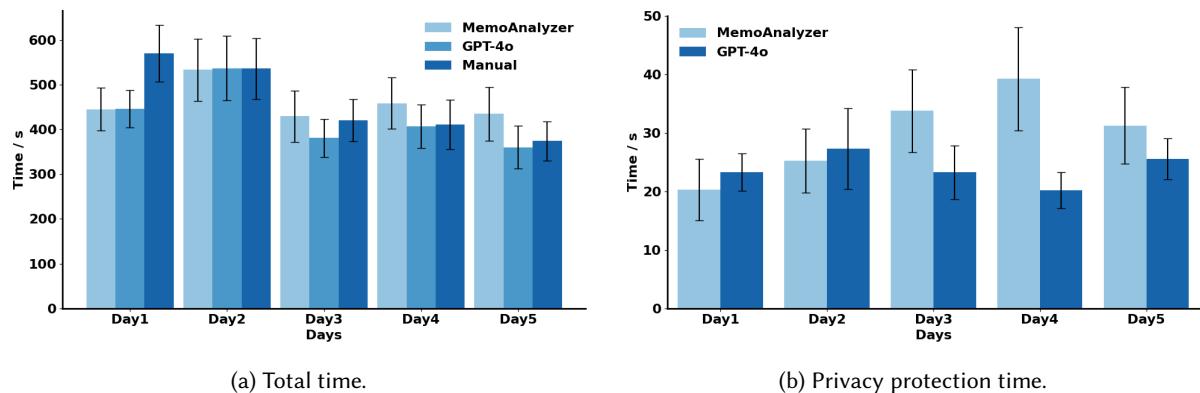| (a) Total time. | (b) Privacy protection time. |

Fig. 6. (a) Total time and (b) Privacy protection time for each day for each technique. Errorbar indicated one standard error.

Over the five days, time had no significant effect on the total completion time for *MemoAnalyzer* ($F_{4,140} = 1.32$, $p = .26$). In contrast, time significantly affected the total completion time for both *GPT-4o* ($F_{4,140} = 4.33$, $p < .01$) and *Manual* ($F_{4,140} = 5.01$, $p < .001$). This suggests that MemoAnalyzer maintained consistent task completion and privacy management times.

5.4.3 Interaction Statistics. We calculated the interaction statistics for each day for each technique, as shown in Table ??.

Notify: Notify refers to the average number of notifications per dialogue. MemoAnalyzer consistently achieved a high notification rate, indicating effective detection of private information requiring user control. The stable notification rate on Days 4 and 5 suggests that MemoAnalyzer effectively managed privacy, with participants disclosing less uncontrolled private information.

Click: Click refers to the average number of clicks per task. The click count initially increased, then decreased over five days, reflecting users' growing need for privacy control as concerns heightened, followed by stabilization

as privacy risks diminished. This trend demonstrates MemoAnalyzer's effectiveness, with privacy management stabilizing by Day 5.

Revise: Revise refers to the average number of revisions per task in MemoAnalyzer. Revision frequency increased over time, stabilizing on the last two days, indicating users' growing need to modify private information, with concerns reaching equilibrium as privacy control improved.

Revise(4o): Revise is defined as the average number of revise times per task for GPT-4o. For GPT-4o, the revision rate was consistently lower than MemoAnalyzer's. Users found it challenging to identify and manage GPT-4o's private information, resulting in fewer proactive deletions and a higher privacy risk.

Use Input: Use Input refers to the average use of past input in each inference. The frequency of past input usage increased as participants incorporated more historical data into their dialogues, leading to greater reliance on past inputs over time, especially as dialogue histories accumulated.

Use Memory: Use Memory refers to the average utilization of memory for each inference. The usage rate remained stable across the five days, which indicates the robustness and effectiveness of MemoAnalyzer.

Coverage: As MemoAnalyzer would highlight the original text when inference, we defined coverage as the overlap of users' memory and the highlighted original text. Coverage remained high throughout the experiment, with a slight dip on Day 2 as participants experimented with modifications beyond the highlighted text, which proved less effective, returning the coverage to 96% by Day 3.

The results demonstrate the effectiveness of MemoAnalyzer across various dimensions. Participants showed consistent modification patterns, typically adjusting their private information after viewing the answers, or during the model's output phase in the next round. This reduced overall task time. Two behaviors were observed: 1) clicking on high-risk privacy items, and 2) reviewing all inferred private information to examine the inference process. Participants primarily modified or deleted highlighted content, confirming MemoAnalyzer's ability to provide precise privacy control.

5.4.4 Subjective Ratings. Figure 8 showed the participants' subjective ratings. We found significant effects of techniques on all dimensions ($p < .05$). Notably, *MemoAnalyzer* was praised as for its higher satisfaction ($\chi_2^2 = 51.6, p < .001$, post-hoc $p < .001$) and perceived control ($\chi_2^2 = 16.3, p < .001$, post-hoc $p < .001$) compared to the manual baseline. This demonstrated the effectiveness of MemoAnalyzer's collaborative design to provide users with control as well as maintaining the machine agency and accuracy in the same time.

MemoAnalyzer were further favored for its superior privacy risk protection effect ($\chi_2^2 = 46.9, p < .001$, post-hoc $p < .001$ compared with *Manual*, $p < .01$ compared with *GPT-4o*) and the effectiveness ($\chi_2^2 = 46.9, p < .001$, post-hoc $p < .001$ compared with *Manual*, $p < .01$ compared with *GPT-4o*) compared with *Manual* and *GPT-4o*. This demonstrates the system's strong ability to protect sensitive data while ensuring usability and speed. It highlights *MemoAnalyzer*'s effectiveness in mitigating privacy risks, surpassing traditional methods, and aligning with our design goals by fostering privacy awareness through intuitive, user-driven control.

In terms of cognitive load and effort, *MemoAnalyzer* reduced physical ($\chi_2^2 = 48.9, p < .001$, post-hoc $p < .001$) and mental demands ($\chi_2^2 = 39.0, p < .001$, post-hoc $p < .001$), frustration ($\chi_2^2 = 16.6, p < .001$, post-hoc $p < .001$), temporal demand ($\chi_2^2 = 18.6, p < .001$, post-hoc $p < .001$), performance ($\chi_2^2 = 29.8, p < .001$, post-hoc $p < .001$) and effort ($\chi_2^2 = 29.9, p < .001$, post-hoc $p < .001$) compared with *Manual*. These proved *MemoAnalyzer* was easier to use and could manage users' privacy easily.

5.4.5 Subjective Comments. **Timely Reminding and reflection:** 29/36 participants commented the notification could help them gain privacy awareness better through visualization of the private information. In particular, they gave high comments for the confidence level and the information sensitivity visualization design. P15 commented that *"I used to first look at the content of the notification, then pay attention to the color of the pop-up box. Brighter colors catch my attention more easily. The confidence percentage and color transparency help me realize how accurate*
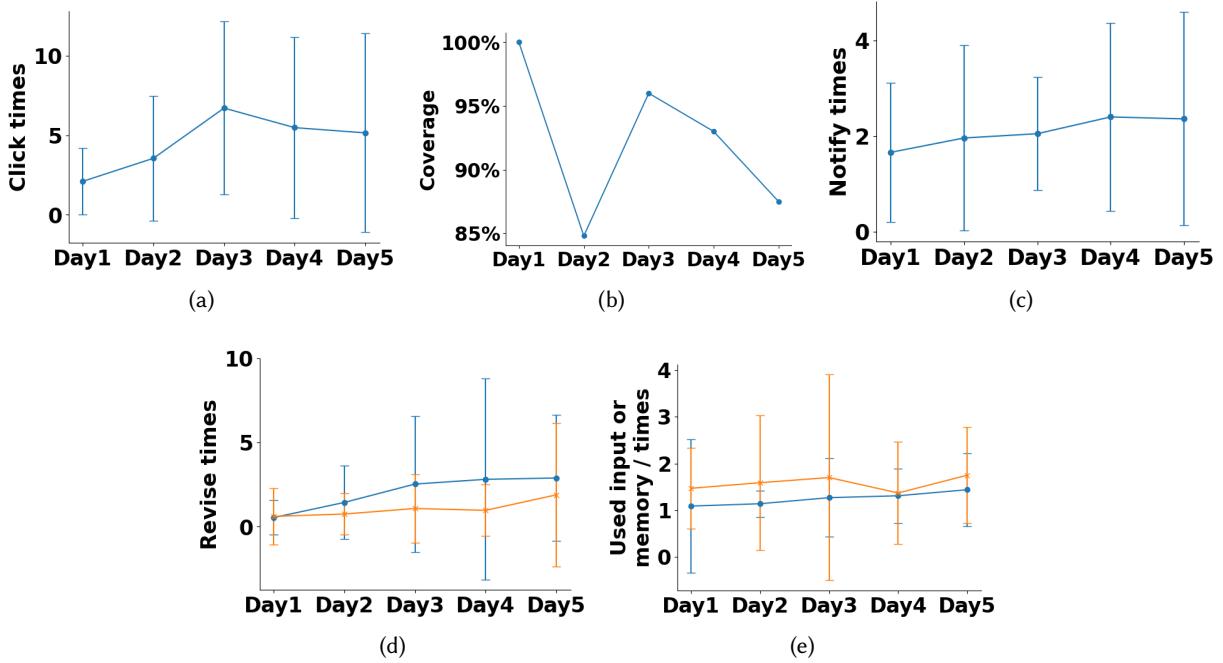
Fig. 7. Trends of different metrics over five days: (a) **Click**: number of user clicks on inferred information, (b) **Coverage**: overlap between users' memory and the highlighted original text, (c) **Notify**: number of notification pop-ups, (d) **Revise**: the orange line shows the number of user modifications or deletions of past inputs/memories for MemoAnalyzer, while the blue line shows deletions of past memories for GPT-4o, (e) **Use of input and memory**: the orange line shows the average number of memories used per inference, and the blue line shows the average number of past inputs used per inference. Errorbar in (a), (c), (d), (e) indicated one standard deviation.

the information is." 28/36 participants found the notification help user reflect, some privacy information user did not notice originally will lead to privacy leaks. In all notifications, 26/36 participants give priority to relevant with themselves and correct personal information.

Increasing Privacy Awareness: We also found the increase of participants' trust towards the MemoAnalyzer(19/36 participants), which echoed the previous literature that memory management [88] and transparent design [68] could increase participants' trust. They thought the information origin and the visualization of the important information also help them modify the private information faster. *"I find this plugin very helpful, as it enables me to identify where privacy may be exposed in my conversations. This allows me to improve my dialogues while managing my privacy more effectively." (P23)* P12 also thought that *"It effectively highlights information I unintentionally reveal that could be inferred by the model."* After 5 days of the experiment, 24/36 participants understood the meaning of private information better than before the experiment.

Participants also gave constructive comments. For example, P25 commented that *"When using these two technologies, I find it difficult to be aware of the need to manage private information most of the time. When reviewing memory in Baseline-GPT, I adopt the same privacy management strategy as in MemoAnalyzer. While using Baseline-Manual, I try to add only content that does not contain private information; however, if certain private details must be entered, I will delete them after completing all tasks.".* P10 also argued that *"The advantage of the*
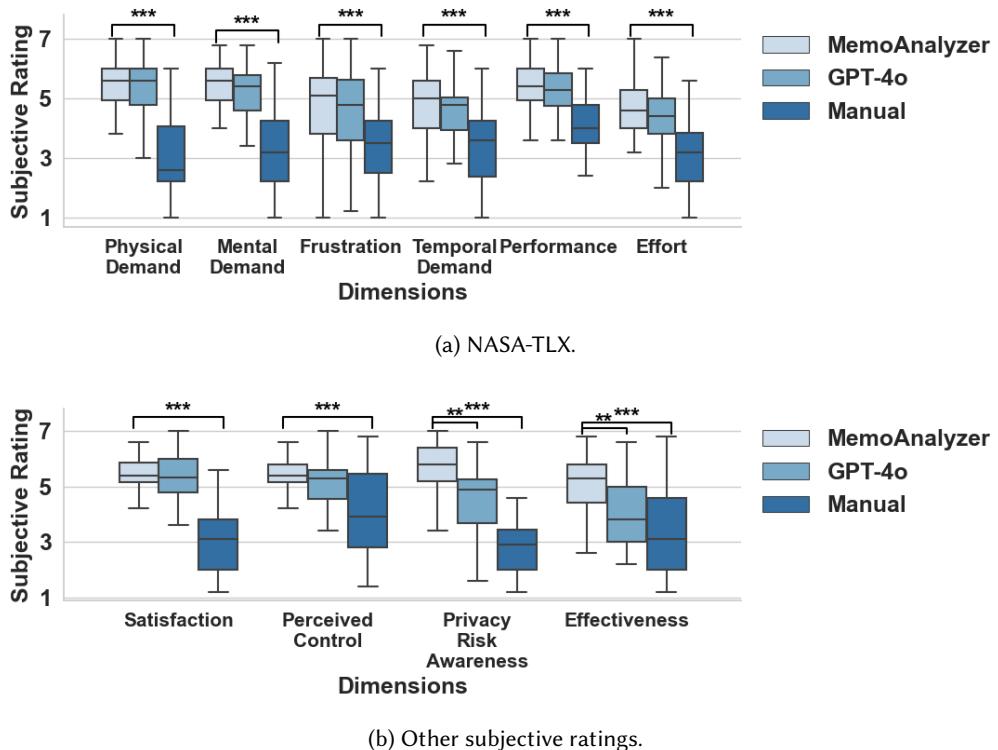
(a) NASA-TLX.

(b) Other subjective ratings.

Fig. 8. The (a) NASA-TLX, and (b) other subjective ratings of all participants (7: most positive, 1: most negative). Errorbar indicated one standard deviation. The significant differences between conditions were from post-hoc analysis.

plugin is that it clearly identifies information I unintentionally reveal, which could be inferred by the model. However, its downside is the low frequency of use in my tasks, which may require extended usage to accumulate more text for inference."

6 DISCUSSION AND DESIGN IMPLICATIONS

6.1 Emerging Issues of Private Memory Information

The emerging privacy risks associated with language models, such as membership inference attacks [76] and data leakage attacks [63], have heightened concerns in the privacy landscape of human-AI interaction. The persistence of memory in LLMs further exacerbates these challenges by intensifying the risk of data leakage. Participants were generally unaware of the privacy risks associated with RAG memory and were unfamiliar with its implications. While they recognized the existence of context memory, they lacked understanding of how to mitigate associated risks, complicating the integration of memory into LLMs. Despite promises from companies like OpenAI to exclude private information unless necessary, participants often disclosed sensitive data that was required for subsequent tasks [93], highlighting the need for effective management. Our work represents the first attempt to address memory-related privacy concerns, leaving place for future exploration in quantitative regulation and algorithmic governance of memory systems [88].

6.2 Designing and Managing Private Information in Memories

Memories are essential in human-AI interactions, enhancing personalization by retaining past exchanges. While the risks of using memory data for training are similar to those of using direct user inputs, the threats are more severe because memory data is retained longer, increasing the potential for misuse. We further found users seldom noticed and cared the existence of memory (see Section 3.5). This paper presents the first framework for managing private information in AI memories, specifically addressing vulnerabilities to membership inference attacks [63] and privacy leakage [76]. Previous research [88] has addressed memory usage in AI systems but has overlooked the private information within memories and failed to model users' mental constructs regarding privacy. Recognizing that not all memory information is essential, we propose a collaborative method involving both users and large language models to select, retain, and utilize memory data according to users' preference [3]. This approach aims to optimize the balance between leveraging valuable memory information and safeguarding user privacy.

6.3 Feasibility of MemoAnalyzer

In this paper we validated the MemoAnalyzer in daily usage cases, where MemoAnalyzer reached better privacy protection effect with comparable time cost. MemoAnalyzer is also expected to be far better given the extreme usage case where private information are ample. We found in this extreme case participants could complete time with a total time comparable with the GPT and the manual baseline. Additionally, participants could control the privacy to 77.7% of the GPT baseline and 95.4% of the manual baseline calculated by number when inferred with GPT-4o. We envisioned MemoAnalyzer as the first system in effectively controlling the memory both no matter in the realistic setting or "cold start" setting (as on Day-1) or in an extreme setting (in Day-3 to Day-4), which could be adapted to various LLMs tasks such as co-writing [89] and ideation [49].

MemoAnalyzer adopted the reactive setting [20] for human-AI collaboration. This enabled participants the proactive control which increased their agency [33] and enhanced their privacy awareness (see Figure 8). Participants were found to proactively determine which information to click and select the private information to control according to their preference (see Section 5.4.3). The reactive design also enabled participants to check before the system automatically delete their important information, which balanced performance enhancement and privacy protection (see Figure 5 and 6).

6.4 Design Implications

We proposed several design implications for the memory management of LLMs to fertilize future design.

Balancing Control and Efficiency in Private Memory Management Through Proactiveness Levels. We used hierarchical design to balance the control and efficiency, where users could click and collapse to view the original source of private information inference result. This allowed the completion of main task while maintaining efficiency for privacy protection. Hierarchical design in privacy protection could also be leveraged in privacy and security settings on smartphones [10] and text anonymization [78], where multiple levels of information or information source needed to be demonstrated efficiently and in the same time reduce users' mental load.

Visualizing the Source of the Privacy Risk to Foster Privacy Awareness. MemoAnalyzer enhance user privacy awareness by visually displaying how inferences are made from stored memories, which similar systems in online chatting [76] and personalization [3] scenarios could also adopt. A clear, intuitive visualization of the data relationships and inferences, perhaps through highlighting key phrases or linking them to past user inputs, will help users understand how their personal information is being processed. This transparency can promote informed decision-making regarding which memories or information to retain, modify, or delete, ultimately fostering greater trust in the system's privacy protections.

Assigning Distinct Roles to Users and AI to Enhance Agency and Efficiency. Assigning users and AI different roles in the privacy management process can significantly improve both perceived agency and efficiency. For instance, users could focus on indicating their preferences or highlighting sensitive information [3], while the AI executes memory adjustments such as deletion, modification, or preservation based on those preferences. This role division reduces the cognitive burden on users, allowing them to concentrate on higher-level privacy decisions while ensuring the system efficiently handles the operational aspects of memory management.

7   ETHICAL CONSIDERATIONS

We acknowledged that our research may have ethical issues. We followed Menlo report [6] and Belmont report [7] in designing the studies and tried our best to avoid ethical concerns. In all studies, we compensated participants according to the local wage standard and told the participants at the beginning of the experiment about the potential benefits and harms. Participants were allowed to quit at any time in the experiment if they felt uncomfortable or for other reasons. Our experiment aimed at solving the privacy policy reading problems through designing applications to facilitate reading. The participants may potentially benefit from reading the privacy policy as they could acquire more information regarding their personal information collection. Besides, all the participants' experimental data was stored on a local device with encryption.

8   LIMITATION AND FUTURE WORKS

We acknowledged that our study have limitations and regard these as future directions. First, although we tried to diversify the background of the participants, we recruited the participants through snowball sampling in the campus, which restricted the age and educational background. University students was a group with relatively higher educational background than the average [8] and may understand the memory mechanism easier. This would further highlight the problems that memory mechanism is opaque and hard to understand. Second, the experiment may face social desirability [17] and recall bias [18] of the participants, during which participants may utter more opinions towards memory mechanism. We selected the most common problems of participants and envisioned these were real problems that needed handling.

9   CONCLUSIONS

This paper presents MemoAnalyzer, a proactive memory management system designed to mitigate privacy risks in human-LLM interactions. Our research highlights the opaque nature of memory mechanisms in current LLMs, which users are largely unaware of. Through a semi-structured interview and a five-day user study, we identified a significant gap in user awareness and control over long-term memory retention. MemoAnalyzer effectively addresses these issues by providing transparency, visualization, and user-driven control over private information, which was validated through improvements in privacy awareness, perceived control, and user satisfaction. This work contributes to the broader discourse on privacy-conscious AI design by demonstrating how user-centric privacy tools can enhance trust and control without impacting system performance. Future work can explore the scalability of MemoAnalyzer in more diverse real-world settings and investigate further improvements in user interaction with privacy management systems.

REFERENCES

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514. https://doi.org/10.1126/science.aaa1465 arXiv:https://www.science.org/doi/pdf/10.1126/science.aaa1465

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[3] 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 CHI Conference on Human Factors in Computing Systems*. 1–21.

[4] Rana Ayyub. 2018. In India, journalists face slut-shaming and rape threats. *New York Times* 22 (2018).

[5] 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.

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

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

[8] Julian R Betts and Darlene Morell. 1999. The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects. *Journal of human Resources* (1999), 268–293.

[9] Jaspreet Bhatia and Travis D Breaux. 2018. Empirical measurement of perceived privacy risk. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 6 (2018), 1–47.

[10] Amel Bourdoucen and Janne Lindqvist. 2024. Privacy of Default Apps in Apple's Mobile Ecosystem. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–32.

[11] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

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

[13] Matt Burgess. 2021. The Biggest Deepfake Abuse Site Is Growing in Disturbing Ways. *Wired* (2021).

[14] Fred H Cate. 2016. The failure of fair information practice principles. In *Consumer Protection in the Age of the'Information Economy'*. Routledge, 341–377.

[15] Yanto Chandra, Liang Shang, Yanto Chandra, and Liang Shang. 2019. Inductive coding. *Qualitative research using R: A systematic approach* (2019), 91–106.

[16] Muhao Chen, Chaowei Xiao, Huan Sun, Lei Li, Leon Derczynski, Animashree Anandkumar, and Fei Wang. 2024. Combating Security and Privacy Issues in the Era of Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*. 8–18.

[17] Janne Chung and Gary S Monroe. 2003. Exploring social desirability bias. *Journal of Business Ethics* 44 (2003), 291–302.

[18] Steven S Coughlin. 1990. Recall bias in epidemiologic studies. *Journal of clinical epidemiology* 43, 1 (1990), 87–91.

[19] Aritra Dasgupta, Min Chen, and Robert Kosara. 2013. Measuring Privacy and Utility in Privacy-Preserving Visualization. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 35–47.

[20] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards Human-centered Proactive Conversational Agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 807–818.

[21] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. http://arxiv.org/abs/1702.08608 cite arxiv:1702.08608.

[22] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 211–223. https://doi.org/10.1145/3172944.3172961

[23] Patricia I Fusch Ph D and Lawrence R Ness. 2015. Are we there yet? Data saturation in qualitative research. (2015).

[24] Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W Malone. 2024. A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–11.

[25] Reza Ghaiumy Anaraky, Yao Li, Hichang Cho, Danny Yuxing Huang, Kaileigh Angela Byrne, Bart Knijnenburg, and Oded Nov. 2024. Personalizing Privacy Protection With Individuals' Regulatory Focus: Would You Preserve or Enhance Your Information Privacy?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[26] Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics* (1961), 148–170.

[27] Victoria Groom and Ryan Calo. 2011. Reversing the privacy paradox: An experimental study. TPRC.

[28] Saul Hansell. 2002. Compressed Data: The Big Yahoo Privacy Storm That Wasn't. *New York Times* (2002), 4.

[29] Julia Hanson, Miranda Wei, Sophie Veys, Matthew Kugler, Lior Strahilevitz, and Blase Ur. 2020. Taking Data Out of Context to Hyper-Personalize Ads: Crowdworkers' Privacy Perceptions and Decisions to Disclose Private Information. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[30] Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2024. HiAgent: Hierarchical Working Memory Management for Solving Long-Horizon Agent Tasks with Large Language Model. *arXiv preprint arXiv:2408.09559* (2024).

[31] 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. 1–3.

[32] 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. https://doi.org/10.1145/3586182.3615796

[33] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2022. Ai in your mind: Counterbalancing perceived agency and experience in human-ai interaction. In *Chi conference on human factors in computing systems extended abstracts*. 1–10.

[34] 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.

[35] Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A Choquette-Choo, and Zheng Xu. 2023. User inference attacks on large language models. *arXiv preprint arXiv:2310.09266* (2023).

[36] Judy Kendall. 1999. Axial coding and the grounded theory controversy. *Western journal of nursing research* 21, 6 (1999), 743–757.

[37] Shahedul Huq Khandkar. 2009. Open coding. *University of Calgary* 23, 2009 (2009), 2009.

[38] Hyunjung Kim and Woohun Lee. 2009. Designing unobtrusive interfaces with minimal presence. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. 3673–3678.

[39] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. ProPILE: Probing Privacy Leakage in Large Language Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 20750–20762. https://proceedings.neurips.cc/paper_files/paper/2023/file/420678bb4c8251ab30e765bc27c3b047-Paper-Conference.pdf

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

[41] Youjeong Kim and S Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior* 28, 1 (2012), 241–250.

[42] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[43] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S Rodriguez, and Jon E Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[44] Sam Levin. 2017. New AI can guess whether you're gay or straight from a photograph. *The Guardian* 8 (2017).

[45] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155* (2016).

[46] Sheng Liang, Mengjie Zhao, and Hinrich Schuetze. 2022. Modular and Parameter-Efficient Multimodal Fusion with Prompting. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2976–2985. https://doi.org/10.18653/v1/2022.findings-acl.234

[47] 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

[48] 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. *arXiv preprint arXiv:2311.08719* (2023).

[49] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.

[50] Yu-li Liu, Wenjia Yan, Bo Hu, Yunya Song, and Zhi Lin. 2023. Chatbots or Humans? The Roles of Agent Identity and Information Sensitivity in Privacy Management and Behavioral Intention: A Comparative Study between China and the United States. In *73rd Annual International Communication Association Conference (ICA 2023): Reclaiming Authenticity in Communication*.

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

[52] Yehong Luo, Xiangjun Ma, and Yuzi Yi. 2024. Subjective Privacy Information: Concepts, Models and Characteristics. In *2024 IEEE 14th International Conference on Electronics Information and Emergency Communication (ICEIEC)*. 94–97. https://doi.org/10.1109/ICEIEC61773.2024.10561732

[53] Zheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability Controllable Biomedical Document Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4667–4680. https://doi.org/10.18653/v1/2022.findings-emnlp.343

[54] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[55] Dan Milmo. 2021. Amazon asks Ring owners to respect privacy after court rules usage broke law. *The Guardian* (2021).

[56] OpenAI. 2023. Memory and New Controls for ChatGPT. https://openai.com/index/custom-instructions-for-chatgpt/ Accessed: 2024-09-12.

[57] OpenAI. 2023. Memory and New Controls for ChatGPT. https://openai.com/index/memory-and-new-controls-for-chatgpt/ Accessed: 2024-09-12.

[58] OpenAI. 2023. Tempary Chats. https://help.openai.com/en/articles/8914046-temporary-chat-faq/ Accessed: 2024-09-12.

[59] OpenAI. 2024. Memory and New Controls for ChatGPT. https://openai.com/blog/memory-and-new-controls-for-chatgpt. Accessed: 2024-08-30.

[60] OpenAI. 2024. Memory and New Controls for ChatGPT. https://openai.com/index/memory-and-new-controls-for-chatgpt/. Accessed: 2024-08-31.

[61] Kathrin Otrel-Cass, Bronwen Cowie, and Michael Maguire. 2010. Taking video cameras into the classroom. (2010).

[62] Saurabh Pahune and Manoj Chandrasekharan. 2023. Several categories of large language models (llms): A short survey. *arXiv preprint arXiv:2307.10188* (2023).

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

[64] Rachel Pannett. 2022. German police used a tracing app to scout crime witnesses. Some fear that's fuel for covid conspiracists. *Washington Post* 13 (2022).

[65] Sharrona Pearl. 2010. *About faces: Physiognomy in nineteenth-century Britain.* Harvard University Press.

[66] Charith Peris, Christophe Dupuy, Jimit Majmudar, Rahil Parikh, Sami Smaili, Richard Zemel, and Rahul Gupta. 2023. Privacy in the time of language models. In *Proceedings of the sixteenth ACM international conference on web search and data mining*. 1291–1292.

[67] Chen Ping, Huangfu Da-Peng, and Luo Zu-Ying. 2018. Automatic attendance face recognition for real classroom environments. In *Proceedings of the 2018 2nd international conference on big data and internet of things*. 65–70.

[68] Daniel Reinhardt, Johannes Borchard, and Jörn Hurtienne. 2021. Visual interactive privacy policy: The better choice?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[69] Grega Repovš and Alan Baddeley. 2006. The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience* 139, 1 (2006), 5–21.

[70] S Sapna, S Sandhya, Ramya D Shetty, Spurthy Maria Pais, and Shrutilipi Bhattacharjee. 2023. YOLOv5 Model-based Ship Detection in High Resolution SAR Images. In *2023 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. 1–6. https://doi.org/10.1109/CONECCT57959.2023.10234764

[71] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118 arXiv:https://doi.org/10.1080/10447318.2020.1741118

[72] T Simonite. 2018. Facebook can now find your face, even when it's not tagged.

[73] Daniel J Solove. 2005. A taxonomy of privacy. *U. Pa. l. Rev.* 154 (2005), 477.

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

[75] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298* (2023).

[76] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=kmn0BhQk7p

[77] Luke Stark and Jevan Hutson. 2021. Physiognomic artificial intelligence. *Fordham Intell. Prop. Media & Ent. LJ* 32 (2021), 922.

[78] Dimitri Staufer, Frank Pallas, and Bettina Berendt. 2024. Silencing the Risk, Not the Whistle: A Semi-automated Text Sanitization Tool for Mitigating the Risk of Whistleblower Re-Identification. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 733–745.

[79] Nili Steinfeld. 2016. "I agree to the terms and conditions":(How) do users read privacy policies online? An eye-tracking experiment. *Computers in human behavior* 55 (2016), 992–1000.

[80] Ari Ezra Waldman. 2018. *Privacy as trust: Information privacy for an information age.* Cambridge University Press.

[81] 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).

[82] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2024. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems* 36 (2024).

[83] Xumeng Wang, Chris Bryan, Yiran Li, Rusheng Pan, Yanling Liu, Wei Chen, and Kwan-Liu Ma. 2022. Umbra: A Visual Analysis Approach for Defense Construction Against Inference Attacks on Sensitive Information. *IEEE Transactions on Visualization and Computer Graphics* 28, 7 (2022), 2776–2790. https://doi.org/10.1109/TVCG.2020.3037670

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

[85] 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.

[86] 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.

[87] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[88] Hen Chen Yen. 2024. *Memolet: Reifying the Reuse of User-AI Conversational Memories*. Master's thesis. University of Waterloo.

[89] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 841–852.

[90] Saralin Zassman, Craig S Kaplan, and Daniel Vogel. 2024. Mindful Scroll: An Infinite Scroll Abstract Colouring App for Mindfulness. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[91] S Zhang. 2018. Personalizing dialogue agents: I have a dog, do you have pets too. *arXiv preprint arXiv:1801.07243* (2018).

[92] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501* (2024).

[93] 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.

[94] 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.

[95] 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).

[96] 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. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 5808–5820. https://doi.org/10.18653/v1/2022.naacl-main.426

[97] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19724–19731.

[98] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing Large Language Models with Long-Term Memory. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 17 (Mar. 2024), 19724–19731. https://doi.org/10.1609/aaai.v38i17.29946

[99] Yeshuang Zhu, Shichao Yue, Chun Yu, and Yuanchun Shi. 2017. CEPT: Collaborative editing tool for non-native authors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 273–285.

[100] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics* 7 (2015), 347–360.

A THE QUESTIONS IN STUDY 1

The followings are the questions in Study 1.

A.1 Demographics

1. What is your age range?

2. What is your highest level of education?

3. What LLMs have you used?

4. How well do you think you understand LLMs products? (self-evaluation, 1: very little, 5: very well)

5. How often do you use LLMs products? (self-evaluation, 1: less than once a year, 2: several times a year, 3: several times a month, 4: serveal times a week, 5: several times a day)

6. How well do you understand AI technology? (self-evaluation, 1: very little, 5: very well)

7. what is your field of study or education focus? (IT-related, privacy and security-related, design-related, others, multiple selections allowed)

A.2 Memroy-related Questions

1. How do you think the memory mechanism of LLMs works?

2. Are you aware that LLMs sometimes remember things you've said before? Can you provide a specific example?

3. What benefits do you think the memory function of LLMs can provide?

4. (Follow-up to the previous question) Have you experienced such benefits? How exactly, and why?

5. What privacy threats do you think the memory function of LLMs might pose? (Please elaborate)

6. (Follow-up to the previous question) Have you experienced such threats? How exactly, and why?

7. Are you willing to use the memory function of LLMs? Why? (You can discuss work or study examples with clear formats)

(Below is instruction) This memory mechanism discussed here works through external knowledge storage and retrieval, essentially following the Retrieval Augmented Generation (RAG) approach. During user input, the LLMs detects whether to store certain information and adds corresponding natural language or vectorized representations to the memory, which enhances the model's capabilities when used. The memory function can offer users a more personalized experience, with the information coming from user input.

8. After this introduction, what privacy threats do you think the memory function or LLMs?

9. How severe do you think these threats are? (1: very low, 5: very high)?

10. How do you think the memory function of LLMs should serve you?

A.3 Inference-related Questions

1. What personal information do you think the following text reveals about you?

2. How high do you think the privacy risk of this information is?

3. If a LLM can infer this kind of personal information, what do you think the following text reveals about you?

4. How high do you think the privacy risk of this information is?

A.4 Thoughts on Memory

1. what are your expectations for managing AI memory? What are you core needs?

2. What features should an ideal AI memory management system have? How individual people control it?

3. Do you think it's important for an AI memory management system to clearly show the source of the memory (reasoning process, source dialogue)? Why?

4. Do you think it's important for an AI memory management system to clearly show how the memory should be used? How should it be presented during use?

A.5 Thoughts on Inferences

1. What are you expectations or thoughts about AI inferring personal information? What are you core needs?

2. Do you need the system to transparently show the inference process related to privacy?

3. Would you need assistance in understanding what kind of information might be inferred?

4. What kind of visualization would you prefer for this?

5. What solutions do you think could address the issue of AI inferring personal information?