Analysis of the Content of ChatGPT's Memory: Types of Information, Security Implications, and User Perception

Conference Paper · June 2025

DOI: 10.1109/MIPRO65660.2025.11131839

6 authors, including:

Nazmus Sadat
Northern Kentucky University
17 PUBLICATIONS **119** CITATIONS
SEE PROFILE

Nicholas Caporusso
Northern Kentucky University
80 PUBLICATIONS **730** CITATIONS
SEE PROFILE

My Doan
IMT School for Advanced Studies Lucca
8 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Vickey Ghimire
Northern Kentucky University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Analysis of the Content of ChatGPT's Memory: Types of Information, Security Implications, and User Perception

Nazmus Sadat*, Nicholas Caporusso*, My (Hami) Doan†, Vickey Ghimire*
Bijay Dhungana*, Jaljala Shrestha Lama*
* School of Computing and Analytics, Northern Kentucky University, Highland Heights, KY 41099, USA
† IMT - Institute for Advanced Studies, Lucca, Italy
sadatm1@nku.edu

***Abstract*—OpenAI's new "memory" feature enables ChatGPT to provide more personalized and relevant interactions by storing user information from the prompts and using it across conversations. While offering improved responses, the memory feature poses privacy and security challenges. This paper reports a three-fold study investigating ChatGPT's memory feature in more detail. First, we utilized the Knowledge-Attitude-Behavior model and distributed a survey to over 135 users to assess their awareness of ChatGPT's memory functionality, attitudes toward privacy implications, and the behavioral changes prompted by perceived risks. Secondly, memory content from over 55 user accounts was analyzed to evaluate the accuracy, relevance, and privacy of the stored data. Finally, we studied the distribution of the stored data across key categories to obtain insights into what kind of information ChatGPT considers relevant and stores. The findings reveal gaps in user understanding of the memory feature, the need for greater transparency, and the challenges of personalizing LLM agents while safeguarding privacy.**

***Keywords*—ChatGPT, memory, cybersecurity, privacy, LLM**

I. Introduction

In the past couple of years, Large Language Models (LLMs) have become the most prominent application of Artificial Intelligence (AI), driving rapid advancements in generative capabilities and attracting wide-scale attention from academia, industry, and the general public. Examples of popular models include the GPT series developed by OpenAI, Google's Gemini, Meta's Llama, and, more recently, High-Flyer's DeepSeek [1], each of which has demonstrated remarkable proficiency in producing coherent, context-aware text across diverse use cases, from natural language processing to creative content generation.

Among these, OpenAI's ChatGPT has profoundly impacted the AI landscape since its release in 2022. Specifically, ChatGPT's user-friendly interface and ability to generate diverse forms of content through simple text prompts have contributed to the widespread adoption of generative AI-based applications [2], [3]. One of the merits of OpenAI is the continuous experimentation and deployment of new features that render generative AI more pervasive in human tasks. For instance, the integration of ChatGPT into Apple's latest smartphones and computers in 2024 will contribute to individuals' growing reliance on AI-driven tools in everyday life. A significant development in ChatGPT's evolution is the introduction of the persistent "memory" feature. Unlike earlier versions, where each conversation started fresh, ChatGPT now retains information across sessions, creating a digital memory that shapes future interactions. This feature enables the model to provide more personalized and contextually relevant responses, enhancing the user experience. However, this new functionality also raises concerns about the types of information being stored, how users perceive and interact with this feature, and security and privacy risks – especially with sensitive data. It also highlights the need for compliance with regulation frameworks such as the General Data Protection Regulation (GDPR) [4].

Given these considerations, it is crucial to investigate the content of ChatGPT's memory, the types of information it deems relevant, and the implications for user privacy and security. This paper reports a comprehensive three-fold study aimed to address these issues. Our work involved 135 users who participated in a survey asking them to share the information stored by ChatGPT's memory feature in their user accounts. We utilized respondents' answers to assess their awareness, attitudes, and behaviors regarding ChatGPT's memory feature using the Knowledge-Attitude-Behavior (KAB) model. In addition, we analyzed the information stored in their accounts to evaluate the performance of ChatGPT's memory algorithm in extracting accurate and relevant information without potentially compromising users' privacy and security. Finally, we explore the distribution of stored data across key categories (e.g., personal interests, education, preferences, contact information, etc.) to gain insights into the nature of the information retained by the system. This article is structured as follows. Section II reviews relevant literature, Section III details our three studies, Section IV presents results and analysis, and Section V discusses future work and conclusions.

II. Related Work

Most existing studies [5]–[7] in the domain have focused on the architectural aspects of LLM memory, including just-in-time retrieval from external sources (e.g.,

context window, Retrieval Augmented Generation) or on the impact of memory on the model's response quality and resource consumption. Zhang et al. [5] reviewed memory modules in LLMs, examining both textual and parametric designs, while Kwon et al. [6] introduced an attention mechanism inspired by virtual memory to improve efficiency in large-scale deployments. Zhong et al. [7] proposed a method that allows models to recall and update past interactions, refining their understanding of users over time.

Woźniak et al. [8] demonstrated that ChatGPT's memory feature can provide up to 20% gain in recommendation performance by extracting information from the user profiles. While the memory feature can help generate more personalized and helpful responses, it also comes with serious security and privacy concerns. Chen et al. [9] warned about risks tied to data retention in personalization tasks. Reliably sanitizing unstructured and context-dependent data can be challenging because automated filters may fail to recognize implicit private details [10]. Moreover, several research works have shown that inaccurate user information stored in memory can reinforce biases related to gender, race, and nationality, worsening existing issues in AI-generated content reinforcement [11]–[13].

Some studies have explored the risks associated with ChatGPT and other LLMs from different perspectives, including launching cyber attacks [14] and privacy and ethical implications [15], [16]. However, to our knowledge, no existing studies have investigated the security and privacy risks of ChatGPT's "memory" function, specifically focusing on the nature and cybersecurity implications of the data retained within this feature. This gap likely stems from the feature's novelty, having been introduced to public users only in September 2024.

III. Study

This section outlines our three-part study on (i) user knowledge, attitudes, and behavior toward ChatGPT's memory, (ii) the security, accuracy, and relevance of stored data, and (iii) the types of information retained.

A. Study 1. Knowledge, Attitude, and Behavior Assessment

The Knowledge, Attitude, and Behavior (KAB) framework is a theoretical model used in behavioral research to explore the relationship between what people know (knowledge), how they feel (attitude), and what they do (behavior) [17]. The framework posits that knowledge shapes attitudes, which in turn influence behavior. By breaking down these interconnected elements, we can better understand and predict human actions in response to a given phenomenon. Initially developed for improving patient healthcare adherence [17], the approach now finds applications across various fields, including healthcare, education, and cybersecurity [18].

In this study, we used the KAB framework to determine the users' knowledge about ChatGPT's memory feature, their behavior toward it, especially their cybersecurity and privacy concerns, and finally, how their behavior changed regarding sharing the memory content and future precautions they could take to be safe from cybersecurity risks.

- *Knowledge:* We examined how many participants were aware of the "memory" feature of ChatGPT and whether they knew what information ChatGPT had stored about them.
- *Attitude:* We explored participants' perspectives on sharing the contents of their ChatGPT memory. Specifically, we wanted to understand their comfort levels with this feature being automatically enabled, as well as their views on the usefulness of the stored content. Additionally, we examined their perceptions of the quality of the content, focusing on aspects such as accuracy, recency, privacy, and relevance. Finally, we investigated the reasons behind their comfort or discomfort with sharing this information.
- *Behavior:* We also examined participants' actual behaviors regarding the sharing of their ChatGPT memory content. Specifically, we wanted to see whether they would follow through with sharing the content after reviewing what was stored, even if they had previously expressed comfort with the idea. We also studied their intended behaviors based on their knowledge of the feature and their attitudes toward it.

For data collection, we developed a 26-question survey featuring a mix of question formats, including multiple-choice options, five-point Likert scales, and open-ended text boxes. The questionnaire was structured to track shifts in participants' self-reported attitudes and behaviors as they gained new insights and knowledge throughout the study. The survey began by assessing participants' knowledge of and attitude toward ChatGPT's memory feature. Next, they were asked to rate the quality of their ChatGPT memory content and indicate whether they were comfortable sharing this content with us. If participants consented, they were then asked to share their memory content. This step allowed us to observe whether participants would actually share their content after initially agreeing to do so. To help those unfamiliar with accessing their memory, we provided instructions on how to retrieve and copy the content.

Following this, participants answered questions about the storage of potentially sensitive information in ChatGPT's memory, including personal identifiers, contact details, health information, political affiliations, etc. They were also asked about their level of concern regarding the feature and what actions they might take after getting a better understanding of it. Additionally, the survey included demographic questions covering occupation, age group, gender, field of study or professional background, familiarity with cybersecurity, frequency of ChatGPT usage, and how often they reviewed or managed their memory.

B. Study 2. Analysis of Security, Accuracy, and Relevance

This study examined the quality and characteristics of information stored in ChatGPT's memory feature through a systematic evaluation of the content from users' accounts.

The research methodology involved an expert assessment of the information, focusing on three key dimensions: security, accuracy, and relevance. The selection of these three aspects was driven by their fundamental importance in assessing any information management systems, particularly in the context of AI-powered applications and, specifically, LLMs. In the context of our study, the dimensions are intended as detailed below.

- *Security* addresses the potential risks associated with the information stored about the user. When analyzing security, the evaluation focused on answering the question: "*To which degree ChatGPT's memory feature balances the utility of stored information with the imperative to protect user privacy?*". To this end, our analysis focused on the presence of any safeguards to protect sensitive user data, including personal information and elements that could be utilized for social engineering or health conditions, with specific regard to compliance with privacy regulations.
- *Accuracy* encompasses the correctness of the stored information with respect to the main objective of ChatGPT's memory feature, that is, storing useful data about the user to personalize future conversations. When analyzing accuracy, the evaluation focused on answering the question: "*to which degree does the information represent the user?*". This includes examining whether the content stored by ChatGPT included information regarding the user.
- *Relevance* refers to how well the stored information aligns with the user's needs. To this end, we examined how ChatGPT's memory feature retains information that meaningfully aligns with the user objectives and contributes to future interactions. The evaluation focused on answering the question: "*to which degree can the information be utilized to enhance future interactions with the user?*". This includes how well specific details, dates, preferences, and other facts shared during conversations could represent the user without distortion or degradation over time.

These dimensions should not be viewed as independent metrics but as layers building upon one another, as shown in Figure 1. Information that cannot be secured adequately should not be stored, regardless of its potential utility or accuracy. If the stored information does not accurately represent the user, its presence in the system becomes potentially harmful rather than helpful, regardless of its potential relevance to future interactions, and, thus, should not be stored. Finally, storage is only justified if it serves a meaningful purpose in enhancing the user experience. Therefore, irrelevant information should not be kept. This also serves as a decision framework for information retention: only information that satisfies all three criteria, that is, being secure, accurate, and relevant, should be maintained in the system's memory.

As ChatGPT's memory feature can store many pieces of information extracted from multiple different conversations, each user's memory content was divided into "memory units" that were evaluated and scored individually. Five experts were involved in the assessment of the dimensions. Each expert analyzed the content from each user's memory, and they independently scored each dimension using a Likert scale with values ranging from very poor (1) to very good (5). Although we could use stricter and more reliable measures, in this phase, we aimed to identify the main areas of improvement for the feature rather than measure them accurately. Moreover, using a Likert scale enabled addressing situations that required a fuzzy approach, like pieces of memory that contained the user's first name or last name only instead of the full name.

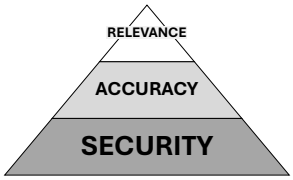

Fig. 1. The framework utilized in Study 2.

C. Study 3. Evaluation of the Information Stored

In our third study, we investigated the types of information that are extracted and stored by ChatGPT's memory feature. This exploration is crucial to understanding the nature of data retention in conversational AI systems, which has significant implications for user privacy, data security, and the ethical use of AI technologies. On the one hand, understanding what information is stored and how it is categorized can help develop more transparent and user-friendly AI systems, ensuring that they align with user expectations and regulatory requirements. On the other hand, as these systems become more pervasive, the need to comprehend their data-handling practices becomes more relevant, especially in closed-source systems such as ChatGPT.

To this end, our methodology combined automated and manual approaches. We defined 15 distinct categories of information that ChatGPT's memory feature could potentially store. These categories were designed to encompass various data types, from personal identifiers (e.g., personal information like the user's name, contact information such as an email address, health information like the presence of conditions, or employment and education data) to contextual conversation details (e.g., details that do not contribute to meaningful profiling). At this preliminary stage of our research, this taxonomy was created based on the potential content of the memory. Then, we utilized an LLM, specifically Phi4, to automatically label each piece of information with one or more predefined categories. Using an LLM for this task allowed us to realize an initial screening and labeling while maintaining a high level of consistency in the labeling process. Following the automated classification, we conducted a manual review of the labels assigned by the LLM to confirm their accuracy.

IV. Results and Discussion

A. Study 1. Knowledge, Attitude, and Behavior Assessment

We collected data from 135 participants, with 63% identifying as male, 34% as female, and 3% as other genders. Participants ranged in age from under 18 to 54. In terms of cybersecurity experience, 67% had little to no experience, 26% had one to two years, 6% had three to five years, and only 1% had more than five years of experience.

To assess familiarity with the memory feature, participants were asked about their ChatGPT usage duration and awareness of the feature. About half of the respondents (48%) were either unaware of it or unsure about its existence (see Figure 2). This is concerning, as ChatGPT may have been utilizing this feature without their knowledge. On the other hand, 51% of participants reported being aware of the feature. This relatively high percentage, given that the feature is pretty new, may be attributed to the background of the sample population, as many participants were from Computer Science or related fields. However, some may have confused the memory feature with ChatGPT's conversation log, potentially inflating the number. Consequently, we anticipate that awareness levels would likely be lower among the general public.

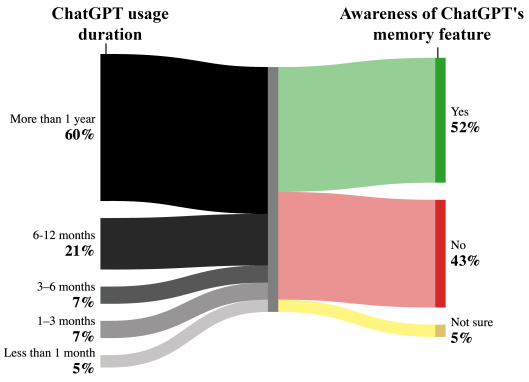

Fig. 2. Usage of ChatGPT vs. awareness of the memory feature.

When asked how comfortable they are about the automatic activation of the memory feature without requiring them to opt-in, participants showed notable differences based on their awareness of this functionality. As shown in Figure 3, 44% of those aware of the auto-activation were comfortable with it, compared to only 19% of those unaware. Additionally, while 27% of the aware group reported being somewhat or very uncomfortable, this number rose to 63% among those unaware. This significant contrast highlights the importance of transparency, as users tend to feel uneasy when they are not informed about the feature.

We also investigated participants' awareness of the memory feature, their knowledge of how to access stored data, and their familiarity with the specific information stored in their accounts. As shown in Figure 4, while 52% of respondents claimed to be aware of the memory feature before the survey, only 22% were confident in their ability to check the data stored in their account. Additionally,

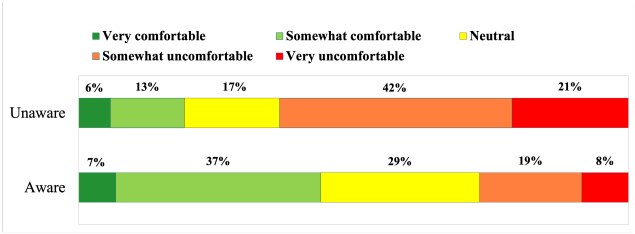

Fig. 3. Knowledge and attitude regarding the automatic activation of the memory feature

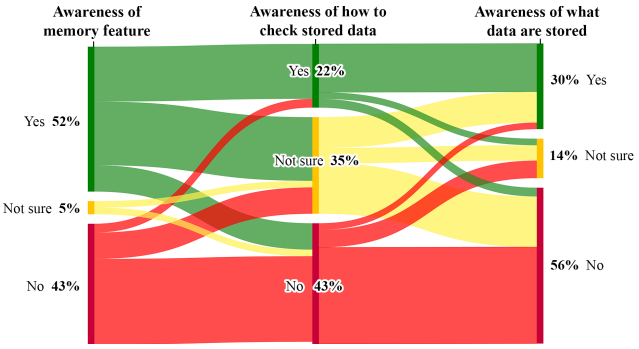

Fig. 4. Knowledge of the memory feature, the process for checking stored data, and the memory content.

when asked if they knew what data was stored in their memory before the survey, only 30% responded affirmatively. These results highlight a significant gap: although some users are broadly aware of the memory feature, most lack detailed knowledge about managing it. The gap in understanding is a critical cybersecurity concern; without a clear grasp of how their information is stored, accessed, or used, users may face risks related to privacy, data accuracy, control, and potential misuse. In addition, given that our participant pool had a more technical background than the general population, it is highly likely that the awareness levels among typical users may be even lower.

The analysis of user attitudes and behaviors toward sharing memory content, as illustrated in Figure 5, reveals interesting relationships between knowledge, comfort levels, and decision-making. When participants were asked to rate their comfort in sharing memory content, 30% expressed some degree of comfort (either very or somewhat comfortable), 32% remained neutral, and 38% reported discomfort (somewhat or very uncomfortable). Notably, prior awareness of the memory feature influenced user attitudes. Among participants who understood how the feature worked prior to the survey, 45% reported discomfort with sharing, whereas only 31% of uninformed users expressed similar hesitancy. This difference suggests that increased awareness of the feature correlates with greater privacy concerns, emphasizing the need for more transparency and clearer user education about how memory functions and the potential security implications.

When prompted to share their ChatGPT memory content, 41% of participants complied, while 37% refused, citing privacy concerns. The remaining 22% did not share

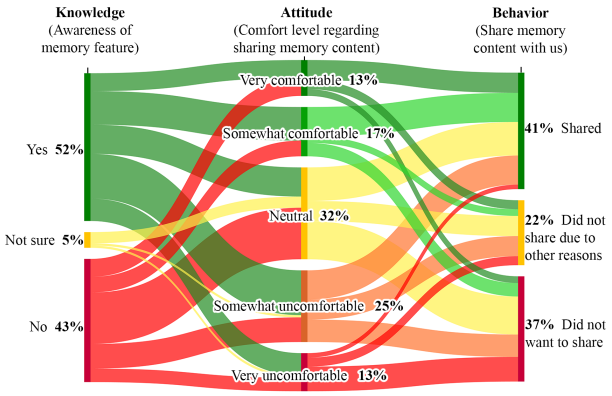

Fig. 5. Knowledge, attitude, and behavior regarding sharing the memory content.

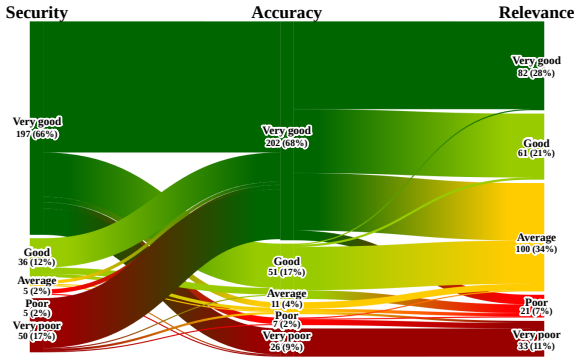

Fig. 6. The results of the analysis of information security, accuracy, and relevance.

due to non-privacy-related reasons, including technical limitations (e.g., difficulties copying content from the ChatGPT iOS app) or because they had intentionally disabled the feature or used ChatGPT without an account.

Interestingly, the data revealed a discrepancy between attitudes and behaviors. Of those who initially reported being uncomfortable or neutral, 58% ultimately chose to share their memory content. Conversely, 19% of those who initially claimed to be comfortable with sharing ultimately declined to share. This behavioral shift suggests two key insights: First, users may share sensitive information unintentionally when unaware of potential privacy risks, emphasizing the ethical responsibility of platforms to educate users. Second, upon reviewing their stored memory content, some users who had initially felt comfortable may have recognized more sensitive information than they anticipated, leading them to withhold sharing. This indicates that privacy concerns are context-dependent and grow with proximity to the data itself.

B. Study 2. Analysis of Security, Accuracy, and Relevance

The results of the analysis, which was conducted in two phases, are summarized in Figure 6. In the first phase, each dimension was considered independent of the others. By doing this, we could determine how much information met each criterion. Our results revealed that, on average, the security of the information scored $4.1 \pm 1.5/5$ (1 being very poor and 5 being very good), suggesting that, in general, ChatGPT's memory feature has a good security level. Accuracy scored the highest value, $4.3 \pm 1.2/5$ on average, confirming that ChatGPT effectively identifies which information represents users' characteristics and preferences. Finally, relevance achieved the lowest score, that is, $3.4 \pm 1.3/5$, which indicates that the content of the memory has the potential to meaningfully contribute to enhancing future interactions.

A more in-depth analysis revealed that 20% of individual pieces of information had a poor or very poor level of security, highlighting potential vulnerabilities in the system's ability to protect sensitive data. While the overall security level is commendable, the presence of poorly secured information raises concerns about compliance with privacy regulations and the potential for misuse. Also, while the system performs well overall, 11% represented the user with poor or very poor accuracy. Indeed, inaccuracies could lead to suboptimal user experiences or even harm if incorrect information is used in critical contexts. Finally, 18% of the information had poor or very poor relevance. The lower relevance score suggests that ChatGPT's memory may not always retain information that meaningfully enhances future interactions, which could limit its ability to provide personalized and contextually appropriate responses.

Subsequently, we analyzed the dimensions through our framework, and we applied each aspect as a progressive filter to determine how much of the information stored in users' accounts met all the criteria. To this end, we considered only information that scored at least four on every dimension. Of 297 individual pieces of information, 233 had good or very good security requirements, and only 66% of the content considered secured also met the accuracy requirements. Finally, only 38% of the information (i.e., 112 pieces) that passed both security and accuracy filters was found relevant for enhancing future interactions.

The findings show that while individual dimensional scores might appear promising when considered separately, ChatGPT's memory feature should not store pieces of information that do not meet all three criteria.

C. Study 3. Evaluation of the Information Stored

As in the previous study, we used the 297 distinct data points of 55 survey respondents. The results (see Figure 7) show that *interests* constituted the most frequently identified category (n=209), followed by *education* (n=101), *employment* (n=89), and *preferences* (n=78). This suggests that a significant portion of user interactions revolved around personal hobbies, academic pursuits, and intellectual engagement. The large quantity of information related to education was mainly because our respondents were primarily college students. In contrast, while less frequent, the presence of sensitive data categories, precisely *personal information* (i.e., 18%), *health* (2%), *financial information* (1%), and *contact information* (1%), raises

significant ethical and privacy concerns. As discussed earlier regarding security, even if these categories constitute a smaller proportion of the overall data, their retention necessitates careful consideration regarding how sensitive user information is handled, stored, and potentially shared or exposed in future interactions. The information associated with *other*, that is, 11%, reflects the non-negligible portion of user data that does not fit neatly into existing classifications because it is highly inaccurate and irrelevant, as shown by the results of Study 2.

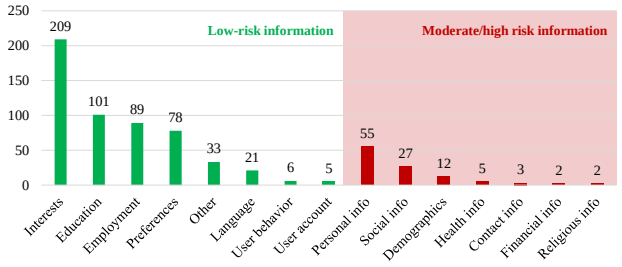

Fig. 7. Distribution of the main types of information stored by ChatGPT memory in users' accounts.

V. Conclusion and Future Work

Our study of the security, accuracy, and relevance of the information stored by ChatGPT's memory in users' accounts shows the presence of poorly secured information, suggesting the need for stronger encryption, access controls, and compliance with privacy regulations. While accuracy scores were high, the memory feature could benefit from mechanisms to verify and update stored information, reducing the risk of misrepresentation. Also, the feature should prioritize retaining information that meaningfully enhances user interactions, potentially through user feedback or adaptive learning algorithms. Finally, our findings from Study 2 highlight the need to give users greater control over what information is stored and how it is used, along with transparency about the system's data retention practices. Our third study, which explored the range of information that ChatGPT's memory feature retains during user interactions, indicates that the system predominantly stores information related to users' interests, education, employment, and personal preferences. However, more sensitive categories were frequently captured, including personal, health, financial, and contact details, reinforcing the findings of Study 1 regarding better data management practices and privacy safeguards. Although one limitation of our study is that the data in a few categories are likely biased due to the student-heavy demographic of the sample, our combined results from Studies 1 and 2 indicate that ChatGPT's memory feature extraction algorithm is generally sensitive in accurately identifying relevant aspects that users find personally and professionally engaging.

As the feature evolves, this preliminary research provides a baseline for future studies to analyze changes in how design and cybersecurity flaws are addressed in future releases by OpenAI. To this end, following ChatGPT memory's rollout plans, we plan to expand our study to a larger population, define a more robust taxonomy for the type of information stored in memory and their associated risks, and develop recommendations for ChatGPT and similar systems.

References

[1] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu *et al.*, "Deepseek llm: Scaling open-source language models with longtermism," *arXiv preprint arXiv:2401.02954*, 2024.

[2] M. Abdullah, A. Madain, and Y. Jararweh, "Chatgpt: Fundamentals, applications and social impacts," in *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. Ieee, 2022, pp. 1–8.

[3] K. Weise, C. Metz, N. Grant, and M. Isaac, "Inside the ai arms race that changed silicon valley forever." *International New York Times*, pp. NA–NA, 2023.

[4] P. Regulation, "Regulation (eu) 2016/679 of the european parliament and of the council," *Regulation (eu)*, vol. 679, p. 2016, 2016.

[5] Z. Zhang, X. Bo, C. Ma, R. Li, X. Chen, Q. Dai, J. Zhu, Z. Dong, and J.-R. Wen, "A survey on the memory mechanism of large language model based agents," *arXiv preprint arXiv:2404.13501*, 2024.

[6] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023, pp. 611–626.

[7] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, "Memorybank: Enhancing large language models with long-term memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 724–19 731.

[8] S. Woźniak, S. Suen, B. Koptyra, M. Kazienko-Sobczuk, A. Szczęsny, P. Kazienko, J. Kocoń, E. Cambria, and K. Kwok, "Explainable and user-controllable profiles including chatgpt-memory: Toward better llm recommendations," *Available at SSRN 4982389*.

[9] J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang *et al.*, "When large language models meet personalization: Perspectives of challenges and opportunities," *World Wide Web*, vol. 27, no. 4, p. 42, 2024.

[10] H. Brown, K. Lee, F. Mireshghallah, R. Shokri, and F. Tramèr, "What does it mean for a language model to preserve privacy?" in *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2022, pp. 2280–2292.

[11] H. Kotek, R. Dockum, and D. Sun, "Gender bias and stereotypes in large language models," in *Proceedings of the ACM collective intelligence conference*, 2023, pp. 12–24.

[12] S. Urchs, V. Thurner, M. Aßenmacher, C. Heumann, and S. Thiemichen, "How prevalent is gender bias in chatgpt?–exploring german and english chatgpt responses," *arXiv preprint arXiv:2310.03031*, 2023.

[13] Z. Talat, A. Névéol, S. Biderman, M. Clinciu, M. Dey, S. Longpre, S. Luccioni, M. Masoud, M. Mitchell, D. Radev *et al.*, "You reap what you sow: On the challenges of bias evaluation under multilingual settings," in *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022, pp. 26–41.

[14] M. Charfeddine, H. M. Kammoun, B. Hamdaoui, and M. Guizani, "Chatgpt's security risks and benefits: offensive and defensive use-cases, mitigation measures, and future implications," *IEEE Access*, 2024.

[15] X. Wu, R. Duan, and J. Ni, "Unveiling security, privacy, and ethical concerns of chatgpt," *Journal of Information and Intelligence*, vol. 2, no. 2, pp. 102–115, 2024.

[16] W. Zhou, X. Zhu, Q.-L. Han, L. Li, X. Chen, S. Wen, and Y. Xiang, "The security of using large language models: A survey with emphasis on chatgpt," *IEEE/CAA Journal of Automatica Sinica*, 2024.

[17] E. P. Bettinghaus, "Health promotion and the knowledge-attitude-behavior continuum," *Preventive medicine*, vol. 15, no. 5, pp. 475–491, 1986.

[18] K. Kannelønning and S. K. Katsikas, "A systematic literature review of how cybersecurity-related behavior has been assessed," *Information & Computer Security*, vol. 31, no. 4, pp. 463–477, 2023.