From Storage to Interpretation: User Perceptions, Practices, and Challenges with Long-term Memory in Agents

Brennan Jones
Faculty of Information
University of Toronto
Toronto, Ontario, Canada
Department of Computing
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China
brennan.jones@utoronto.ca

Nazar Ponochevnyi
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
nazar@cs.toronto.edu

Kelsey Stemmler
Faculty of Information
University of Toronto
Toronto, Ontario, Canada
kelsey@stemmler.tech

Emily Su
Faculty of Information
University of Toronto
Toronto, Ontario, Canada
em.su@mail.utoronto.ca

Young-Ho Kim
NAVER AI Lab
Seongnam, Republic of Korea
yghokim@younghokim.net

Anastasia Kuzminykh
Faculty of Information
University of Toronto
Toronto, Ontario, Canada
anastasia.kuzminykh@utoronto.ca

Abstract

To provide long-term personalized assistance to users, AI agents must have effective long-term memory (LTM). However, there is little understanding of users' perceptions, practices, and challenges with LTM in agents. We interviewed 21 users of agents such as ChatGPT and Claude to understand people's everyday experiences with agent LTM. Our findings shed light on the flow of memory in agents as a three-stage process consisting of (1) information intake, (2) storage and management, and (3) retrieval and interpretation. Users' perceptions of agent LTM are mainly influenced by Stage 3, and thus users' interactions with agent LTM are mainly attempts at influencing and understanding how the agent retrieves and interprets information from memory. Therefore, we recommend that technological approaches to user interaction with agent LTM focus at least as much on memory retrieval and interpretation as they do on memory intake, storage, and management.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

human-AI interaction, human-agent interaction, memory, personalization, language models

ACM Reference Format:
Brennan Jones, Nazar Ponochevnyi, Kelsey Stemmler, Emily Su, Young-Ho Kim, and Anastasia Kuzminykh. 2025. From Storage to Interpretation: User Perceptions, Practices, and Challenges with Long-term Memory in Agents. In *13th International Conference on Human-Agent Interaction (HAI '25), November 10–13, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3765766.3765843

HAI '25, Yokohama, Japan

ACM ISBN 979-8-4007-2178-6/25/11
https://doi.org/10.1145/3765766.3765843

1 Introduction

People are increasingly interacting with AI agents (e.g., ChatGPT) over long periods of time (e.g., [1, 2, 4–6, 8–18, 20–22]). Such agents can provide always-available personalized and contextualized assistance to users. However, in order to achieve this, the user and the agent can benefit from achieving co-understanding of each other, and this relies on the user having proper understanding and influence of the agent's *long-term memory* (LTM). While memory in agents has been extensively covered by prior research (e.g., [7, 23]), there is a lack of understanding of users' perceptions, practices, and challenges with agent LTM. Understanding these could help designers build agent LTM mechanisms that involve users and consider their mental models (in line with [19]).

2 Method

We conducted semi-structured one-on-one interviews on Zoom with 21 participants (each identified as *'P#'*) who use personalized AI tools with LTM on a regular basis. The goal was to better understanding users' expectations of AI systems' LTM mechanisms and their practices of managing these systems' memories over longer periods of time. Each interview lasted between 30 and 90 minutes (50-60 minutes on average), and each participant was reimbursed with $20 CAD. Participants were recruited through social media advertisements, email recruiting, postings in university and academic Slack channels, and snowball sampling. The interviews were recorded and transcribed, and we conducted thematic analysis of the data using an inductive coding approach [3].

3 Findings

Our findings shed light on the flow of memory in agents as a three-stage process, as illustrated in Figure 1. Users' perceptions of agent LTM are mainly influenced by Stage 3, and thus users' interactions with LTM are primarily attempts at influencing or understanding how the agent retrieves and interprets information from LTM.

Users' Perceptions. Users' perceptions of how the agent operates in Stage 2 (*storage and management*) are mostly clear—participants

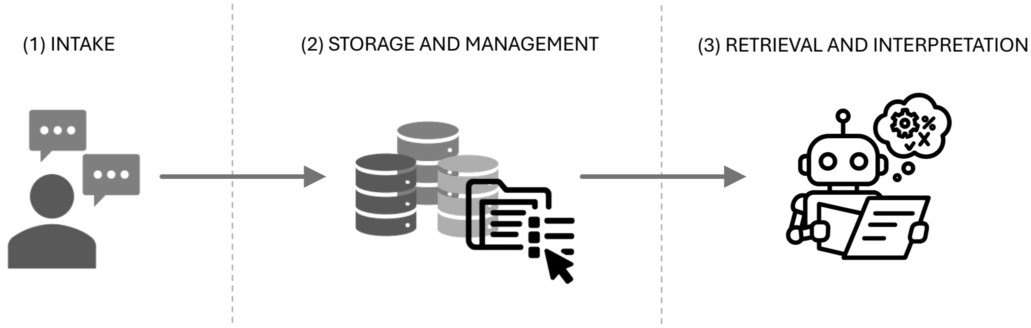

Figure 1: The process of memory flow for an agent with LTM. (1) Discrete data (e.g., messages) come in as input from the user ('Intake'). (2) Some of this discrete data is stored into the agent's LTM ('Storage and Management'). (3) In the future, this information is retrieved from LTM and interpreted by the agent in a new context ('Retrieval and Interpretation').

tended to assume that the agent is either storing all or no information. Tools like ChatGPT save users' chat histories to their accounts, and these are accessible to the user to view or even for continuing previous old chats if desired. This gives some users (e.g., P14) the impression that the system is storing all information that comes in.

Users' perceptions of what the agent does in Stage 3 (*retrieval and interpretation*), however, are less clear. When the agent fails to provide full transparency on how it is retrieving, interpreting, and using information from LTM, users tend to develop their own assumptions about how the agent is doing this. For example, most participants (P1-P6, P8, P13, P15-P20) assumed that if the interface provided the ability to create multiple chat threads (as ChatGPT does), then the agent was able to retrieve memories from within the current chat better than information across chat threads.

Users' Practices of Memory Management. Participants expressed taking a variety of approaches to managing agent LTM. For instance, several participants (P1-P6, P13, P15-P18, P20, P21) shared that they manage conversations about different topics or tasks into different chat sessions, so that the agent would have easier access to memories that are relevant to each specific task. Some participants (P1, P2, P4, P6, P15, P16) expressed that they would copy and paste instructions or relevant context information from previous chat sessions into new chat sessions whenever they thought this was necessary, in order to selectively preserve some of the agent's desired behaviours or memories from previous sessions inside the new session. Others (P8, P9, P14, P19) described using a single chat session for everything, to *"preserve [the] continuity"* (P8, P9) of the conversation or because the user thought that it would be beneficial for the agent to have access to all of its past memories no matter the topic (P14, P19)—for example, to *"see the relationship"* (P19) between multiple topics discussed with the agent.

Participants also used other techniques to manage agent LTM, including: deleting or revising previous chat messages (P18); using 'temporary chats' to prevent information from being saved (P4, P18); repeating or emphasizing certain information or instructions in order to get the agent to remember these better (P4-P6, P9, P12, P14, P19); and asking the agent to summarize what it knows, in case the user is unsure (P16-P18). However, while on the surface level these actions affect the storage and management of LTM (Stage 2), participants expressed that they take these approaches with the goal of influencing when and how the agent retrieves memories and how it interprets them (Stage 3). For instance, most participants expressed that they try to abstract the agent's access to information in its LTM based on the current context of the conversation, such as the current task, project, or domain (e.g., by separating memories related to writing from those related to health advice).

Users' Challenges and Unsatisfied Needs. Our participants' practices with memory management reflect their desires for agency over how and when the agent retrieves and interprets past memories in the present moment, and to have greater knowledge of what prior memories the agent is recalling and how it is using those memories in the moment. However, despite users' practices, many participants still faced challenges and expressed needs that were unsatisfied by current systems.

For example, agents sometimes misinterpret the intent of information in memory. P17 described an instance when she was asking ChatGPT to refine a recipe, and asked it to *"make sure the salt is balanced."* ChatGPT interpreted this as a longer-term preference for all recipes moving forward, rather than just a preference for that particular recipe. P18 expressed that she has had to abandon and delete chat threads because the agent started to misinterpret old prompts, and she had the perception that its behaviours were becoming too restricted over time. In other instances, the agent expresses that it remembers certain information but it does not interpret or use that information to the user's desire. For example, P17 used ChatGPT for coding assistance, and she told ChatGPT that she is using Unity version 6, which the agent then expressed that it would remember moving forward. However, it still kept giving the user guidance that was relevant to previous versions of Unity, reflecting that the agent did not know when that piece of knowledge was useful in the current context.

4 Conclusion

Our findings reveal that users' perceptions of agent LTM are mainly influenced by agents' behaviours, which are affected by how agents *retrieve* and *interpret* (Figure 1, Stage 3) information from LTM. Thus, users' practices with managing and exploring LTM are mainly an attempt to influence and understand how the agent retrieves and interprets information from LTM.

Our insights are based on users' recollections from their everyday usage of agents. However, more insights can be gained by developing, exploring, and comparing approaches to user interaction with LTM, and therefore we recommend this as future work.

References

[1] [n. d.]. Introducing Pi, Your Personal AI. https://inflection.ai/blog/pi

[2] Sabina Asensio-Cuesta, Vicent Blanes-Selva, Manuel Portolés, J Alberto Conejero, and Juan M García-Gómez. 2021. How the Wakamola chatbot studied a university community's lifestyle during the COVID-19 confinement. *Health Informatics Journal* 27, 2 (April 2021), 14604582211017944. doi:10.1177/14604582211017944 Publisher: SAGE Publications Ltd.

[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[4] Aranka Dol, Christina Bode, Hugo Velthuijsen, Tatjana van Strien, and Lisette van Gemert-Pijnen. 2021. Application of three different coaching strategies through a virtual coach for people with emotional eating: a vignette study. *Journal of Eating Disorders* 9, 1 (Jan. 2021), 13. doi:10.1186/s40337-020-00367-4

[5] Ahmed Fadhil and Silvia Gabrielli. 2017. Addressing challenges in promoting healthy lifestyles: the al-chatbot approach. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 261–265. doi:10.1145/3154862.3154914

[6] Paula M. Gardiner, Kelly D. McCue, Lily M. Negash, Teresa Cheng, Laura F. White, Leanne Yinusa-Nyahkoon, Brian W. Jack, and Timothy W. Bickmore. 2017. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: A feasibility randomized control trial. *Patient Education and Counseling* 100, 9 (Sept. 2017), 1720–1729. doi:10.1016/j.pec.2017.04.015

[7] Zihong He, Weizhe Lin, Hao Zheng, Fan Zhang, Matt Jones, Laurence Aitchison, Xuhai Xu, Miao Liu, Per Ola Kristensson, and Junxiao Shen. 2024. Human-inspired Perspectives: A Survey on AI Long-term Memory. doi:10.48550/arXiv.2411.00489 arXiv:2411.00489.

[8] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. WeightMentor, bespoke chatbot for weight loss maintenance: Needs assessment & Development. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2845–2851. doi:10.1109/BIBM47256.2019.8983073

[9] Chin-Yuan Huang, Ming-Chin Yang, Chin-Yu Huang, Yu-Jui Chen, Meng-Lin Wu, and Kai-Wen Chen. 2018. A Chatbot-supported Smart Wireless Interactive Healthcare System for Weight Control and Health Promotion. In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. 1791–1795. doi:10.1109/IEEM.2018.8607399 ISSN: 2157-362X.

[10] Brennan Jones, Yan Xu, Qisheng Li, and Stefan Scherer. 2024. Designing a Proactive Context-Aware AI Chatbot for People's Long-Term Goals. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3613905.3650912

[11] Matthew Jörke, Shardul Sapkota, Lyndsea Warkenthien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James Landay. 2024. Supporting Physical Activity Behavior Change with LLM-Based Conversational Agents. doi:10.48550/arXiv.2405.06061 arXiv:2405.06061 [cs].

[12] Tobias Kowatsch, Kim-Morgaine Lohse, Valérie Erb, Leo Schittenhelm, Helen Galliker, Rea Lehner, and Elaine M. Huang. 2021. Hybrid Ubiquitous Coaching With a Novel Combination of Mobile and Holographic Conversational Agents Targeting Adherence to Home Exercises: Four Design and Evaluation Studies. *Journal of Medical Internet Research* 23, 2 (Feb. 2021), e23612. doi:10.2196/23612 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[13] Dillys Larbi, Kerstin Denecke, and Elia Gabarron. 2022. Usability Testing of a Social Media Chatbot for Increasing Physical Activity Behavior. *Journal of Personalized Medicine* 12, 5 (May 2022), 828. doi:10.3390/jpm12050828 Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[14] Gun Woo (Warren) Park, Payod Panda, Lev Tankelevitch, and Sean Rintel. 2024. CoExplorer: Generative AI Powered 2D and 3D Adaptive Interfaces to Support Intentionality in Video Meetings. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3613905.3650797

[15] Gun Woo Warren Park, Payod Panda, Lev Tankelevitch, and Sean Rintel. 2024. The CoExplorer Technology Probe: A Generative AI-Powered Adaptive Interface to Support Intentionality in Planning and Running Video Meetings. doi:10.1145/3643834.3661507 arXiv:2405.18239 [cs].

[16] SoHyun Park, Anja Thieme, Jeongyun Han, Sungwoo Lee, Wonjong Rhee, and Bongwon Suh. 2021. "I wrote as if I were telling a story to someone I knew.": Designing Chatbot Interactions for Expressive Writing in Mental Health. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 926–941. doi:10.1145/3461778.3462143

[17] Meihua Piao, Jeongeun Kim, Hyeongju Ryu, and Hyeongsuk Lee. 2020. Development and Usability Evaluation of a Healthy Lifestyle Coaching Chatbot Using a Habit Formation Model. *Healthcare Informatics Research* 26, 4 (Oct. 2020), 255–264. doi:10.4258/hir.2020.26.4.255 Publisher: Korean Society of Medical Informatics.

[18] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3411764.3445615

[19] Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3613904.3642466

[20] Yanxin Wu, Ryan Donovan, Binh Vu, Felix Engel, Matthias L. Hemmje, and Haithem Afli. 2020. Chatbot Based Behaviour Analysis for Obesity Support Platform.. In *CERC*. 112–124. https://ceur-ws.org/Vol-2815/CERC2020_paper07.pdf

[21] Yan Xu, Brennan Jones, Hannah Nguyen, Qisheng Li, and Stefan Scherer. 2025. From Goals to Actions: Designing Context-aware LLM Chatbots for New Year Resolutions. In *Proceedings of the 2025 ACM Conversational User Interfaces Conference (CUI 2025) (CUI '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3719160.3736637

[22] Amanda L. Zaleski, Rachel Berkowsky, Kelly Jean Thomas Craig, and Linda S. Pescatello. 2024. Comprehensiveness, Accuracy, and Readability of Exercise Recommendations Provided by an AI-Based Chatbot: Mixed Methods Study. *JMIR Medical Education* 10, 1 (Jan. 2024), e51308. doi:10.2196/51308 Company: JMIR Medical Education Distributor: JMIR Medical Education Institution: JMIR Medical Education Label: JMIR Medical Education Publisher: JMIR Publications Inc., Toronto, Canada.

[23] 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. doi:10.48550/arXiv.2404.13501 arXiv:2404.13501 [cs].