Users' Expectations and Practices with Agent Memory

Brennan Jones
Faculty of Information
University of Toronto
Toronto, Ontario, Canada
brennan.jones@utoronto.ca

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, Gyeonggi, Republic of Korea
yghokim@younghokim.net

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

Abstract

AI agents have the potential to provide long-term personalized assistance to users, and this relies on effective long-term memory. While memory in agents has been extensively covered by prior work, there is little understanding of users' expectations and practices with agent memory. As a preliminary investigation, we interviewed people who use AI tools with memory and analyzed online discussion posts of people's experiences with such tools. We found that users often have incomplete mental models of how agents remember and recall information, and how their memories affect their behaviours. Users generally consider agents' memories as belonging to different categories along a hierarchy from more generalized knowledge to more specific knowledge about the user or task. Users often desire the system's memories to be cleanly organized by these categories. These findings reveal opportunities to design agent memory mechanisms to organize and control access to memories based on users' task-based needs.

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, Kelsey Stemmler, Emily Su, Young-Ho Kim, and Anastasia Kuzminykh. 2025. Users' Expectations and Practices with Agent Memory. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3706599.3720158

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720158

1 Introduction

AI agents, including LLM-based chatbots and tools like ChatGPT, Claude, and Gemini, are becoming more prevalent in people's daily lives. People are also increasingly interacting with such agents over long periods of time—e.g., using them as personal AI assistants (e.g., [2]) and 'coaches' for wellness [7, 8, 10, 12, 14, 25, 26, 31], exercise [3, 17, 19, 20, 35], life goals [16], and work productivity [22, 23, 27].

Such AI agents have the potential to provide always-available personalized and contextualized assistance to users, and to work in collaborative 'complementarity' [30] with the user. However, in order to achieve this, the user and the agent can benefit from achieving co-understanding of each other. One aspect that contributes to co-understanding is *memory*. To achieve sufficient co-understanding in longer-term interactions between humans and agents, the agent needs to remember the right things about the user and the interactions over time, and refer to those memories optimally in giving effective assistance to the user. Further, it is also helpful for the user to have sufficient understanding of the agent's decision-making processes and how its memories and knowledge impact its behaviours.

Memory in LLM-based agents has been extensively covered by prior research (e.g., [11, 36]). Strategies for maintaining an agent's memory include summarization (e.g., [29, 32]) and refinement (e.g., [37]) of conversational and interactional histories, maintaining vector databases and retrieving relevant conversations or information through semantic or linguistic matching (e.g., [13, 24, 38]), maintaining key-value pairs in traditional databases (e.g., [21]), or maintaining lists of 'facts' or 'observations' learned from user interactions (e.g., [1, 5, 24, 33]). There are also systems that allow users to influence or modify agents' memories through, for example, direct manipulation (e.g., [15, 34]). Some systems' memory mechanisms (e.g., [18]) are designed to mimic existing psychological models of human memory (e.g., [4]), while others (e.g., [21]) are designed to mimic existing technological metaphors such as traditional operating system memory. However, there is a lack of understanding of users' expectations of AI systems' memory mechanisms while using such systems over longer periods of time. Furthermore, there is a lack of understanding of users' practices of managing these

systems' memories. Achieving such understanding could help designers build memory mechanisms for agents that involve users, consider their mental models (in line with [30]), and keep the human in-the-loop.

To achieve this understanding, we conducted interviews with six participants who use personalized AI tools with long-term memory on a regular basis, and thematically coded these interviews to explore users' expectations and mental models of such systems' memory mechanisms and their practices with managing memory in these systems. We also conducted a thematic analysis of publicly available posts and discussion threads on Reddit, in which people discussed their experiences using the memory features of personalized AI systems.

Our findings illustrate that, in general, users often have an incomplete understanding of how existing agents remember and recall information from users' interactions. Users generally think of the system's memories as belonging to different categories that can be organized along a hierarchy from more generalized knowledge (e.g., *factual memories* of general knowledge facts, usually contained in the model's training data or fetched externally through retrieval-augmented generation, or RAG [9]), to knowledge about the user (e.g., their personal or social details, preferences, etc.), to more specific knowledge about the domains, projects, or tasks that the agent assists the user with. Users' needs for the system to store and retrieve memories along the different positions of this hierarchy evolve depending on the task or stage of the activity they are working on. Our findings also reveal users' existing approaches to organizing, separating, and abstracting access to memories across different tasks, projects, and domains, including their use and organization of chat threads, customs prompts and instructions for defining personas and roles, fine-tuning models, and using different user accounts. These findings reveal opportunities to design agent memory mechanisms to organize and control access to memories via hierarchical structures based on users' task-based needs.

2 Method

We conducted a study with the goal of better understanding users' expectations of AI systems' memory mechanisms and their practices of managing these systems' memories while using them over longer periods of time. Our data was collected from two sources: (1) one-on-one interviews with participants who use personalized AI tools with memory, and (2) online posts and discussions threads on Reddit, where people discussed their experiences using personalized AI tools with memory. Our research questions were as follows:

(1) **What are people's mental models and expectations of what AI memory systems do?** For example, what does the user expect the AI system to remember? How does the user expect the AI to behave when recalling old memories? How do people expect their information to be abstracted when stored in memory?
(2) **What are people's unsatisfied needs with AI memory systems?** How do these unsatisfied needs affect users' interactions with and perceptions of the AI system?
(3) **What are people's current practices with managing and exploring the memory of a personalized AI system?** For example, how does the user (try to) instruct or get the AI to remember, recall, or forget something? When and why does the user want to manage the system's memories?

We conducted semi-structured one-on-one interviews (see Appendix A for the interview guide) with six participants (each identified as *'P#'* in the findings) who use personalized AI tools with long-term memory on a regular basis. Each interview lasted between 30 and 45 minutes. The interviews were conducted on Zoom, and each participant was reimbursed with CAD$20. The interviews were recorded and transcribed for analysis. Participants also filled out a brief survey (see Appendix B) where they provided basic demographic information and answered questions about their current and prior use of AI tools. Four of our participants identified as female and two as male, ranging from 24 to 35 years old (M = 29, SD = 5). Three participants expressed that they use LLMs or personalized AI tools at least once per day, while the other three expressed that they use these tools at least once per week. Of the specific tools that participants mentioned using on a regular basis, all of our participants expressed using ChatGPT, while three also use Claude, two also use the Microsoft Copilot chatbot, and one each also use Perplexity, NotebookLM, Grammarly's AI features, and the Cursor AI code editor for code completion and chat. Our participants expressed using these tools for activities such as *"productivity tasks, writing assistance, usability analysis," "analyzing qualitative data," "email writing and editing," "code-related questions, brainstorming, error troubleshooting," "learning something from scratch," "conversational [web] search[ing],"* and *"as a personal tutor, search engine, and life companion."* All six participants rated their "level of expertise with using these tools" as 4/5 (with 1 = 'beginner/novice' and 5 = 'advanced/expert'). However, when asked about their "level of *technical knowledge* of LLMs" (i.e., their "technical understanding of how LLMs work" and their "technical understanding of memory in LLMs and LLM-based tools and agents"), four participants rated themselves as 4/5, and one participant each rated themselves as 2/5 and 1/5 (with 1 = 'beginner/novice' and 5 = 'advanced/expert').

We also collected and performed a preliminary analysis of 54 publicly available discussion threads on Reddit, in which people discussed their experiences using the memory features of AI systems. The threads were collected systematically from Reddit communities centred around topics such as AI tools, agents, LLMs, generative AI, and machine learning. In addition, relevant threads were collected from general Q&A Reddit communities where people asked questions related to the memory of AI systems. Discussion threads were collected from these communities using search terms and keywords such as *"memory," "forgetting," "remembering," "chat history," "summarization," "temporary chat,"* and *"personalization."* Please see Appendix C for a list of the Reddit communities (*"subreddits"*) where the discussion threads were collected from.

We conducted a thematic analysis of the interview data and Reddit discussions together, using an inductive coding approach [6]. Codes were developed to address our research questions, then iteratively discussed, refined, and categorized into themes. For example, codes like *"user is surprised about how much the system remembers about them," "user wants the system's memories to be organized or disjoint by project, task, or domain,"* and *"user organizes memories*

by chat threads" were created, representing users' perceptions, unsatisfied needs, and current practices with agent memory. During the categorization phase, we saw themes emerge around desires for system transparency, heterogeneity of the system's memories, temporal dependency of memory needs, and user practices of organizing and separating the system's memories. The first three authors coded the Reddit discussion threads independently, but met to discuss their findings and iterate on the coding on a regular basis. The first author completed most of the coding of the interview data, but discussed, reviewed, and iterated on the codes with the second and third authors.

3 Findings

3.1 Users' Perceptions and Desires for Transparency

Overall, we found that users often have an incomplete understanding of AI systems' memory mechanisms, including what the system remembers, how the system decides what information to store in its memory, and how the system's memories influence its behaviours or outputs. As a result, our participants expressed their desire for the system to provide greater transparency of what it remembers about the user and its previous and ongoing interactions with them.

> *"When I'm conversing in it with some other ways, or, you know, just regular email writing or something, it'll also sometimes say 'memory updated'. And then, but, I think* ***I don't really know how it decides what is important information to remember versus what is not important information to remember*** *in, like the prompts that I write."* — P3

> *"Definitely not [transparent]. Like, I'm... I can never say for sure whether it remembers this thing or not. And even sometimes I tell it to remember, it may forget."* — P6

Users also want the system to provide greater transparency of how its past memories influence its present behaviours:

> *"It would be interesting, like I mentioned,* ***to know which parts of the memory it has about me go into, like, which responses****, and to understand more about how it decides which portions [...] of that list [of memories] is like... you know, [it is] 100 or 200 items... like, that's a lot of things to remember about me. And to decide, like, which parts of that list actually go into a response versus not."* — P3

Some of our participants were surprised about how much the systems they use remember. For instance, our participants who have used ChatGPT, Copilot, and other LLM-based chatbots expressed that they were surprised about how much the system remembers sensitive or personal information about them.

> *"I didn't notice it, but now I'm looking at the memory, it's definitely collected a bunch of information about me. [...] It's scary."* — P1

> *"Yeah, it remembers quite a lot. Oh, my God."* — P2

Participants were particularly surprised that the systems they use preserve memories containing details that they consider to be 'small', 'mundane', or not directly related to the activities that they use them for.

> *"So it saved my supervisor['s name]. Which makes sense. Because it read emails, like, I polish my emails to my supervisor with GPT, so it knows.* ***And the really scary one is, it has my student ID. And then my employee ID.*** *I think I don't put [those] there a lot. Maybe it's just... for the one or two times I'm trying to contact the department, like for support, where I put my information there [in the email], and so I'm really surprised. Like, I didn't keep putting it there, but [ChatGPT's memory] captured it."* — P6

In some cases, the systems were able to use these small details to make detailed inferences about the user—for example, about their personality.

> *"I actually tried asking ChatGPT to guess my MBTI [Myers–Briggs Type Indicator personality test]. And [...] it got three out of the four [indicators] correct. So like that,* ***that kind of threw me off, because I don't talk much about my personality.*** *Like I mentioned, I'm using [ChatGPT] mostly for productivity. I use it to help me, you know, write papers, and then, to, you know, improve [the] language of emails, and things like that. [...] And then ChatGPT guessed [my personality], just based on regular interactions that were mostly about work and research. [...] So I think* ***that was a little, where I was caught off guard.***" — P3

While some were initially amused by this, most participants did not want the system to remember such personal details over a longer term. Therefore, in addition to desiring more awareness and transparency of what/how the system remembers and how its memories influence its behaviours, participants wanted to also have control over these processes. In some cases, participants were okay with the system remembering certain information about them as long as it asked them for their consent.

> *"If it explicitly asks me for basic [personal] information, and then it's up to me whether or not I give it that information, that's perfectly fine."* — P4

Some members of the Reddit community were surprised and impressed by how much the agent remembered about them, expressing that their conversations were improving as a result, and the agent was developing a deeper understanding of them. One individual even felt a social connection with the agent. While this is the case, some others were frustrated at the amount of irrelevant information the agents remembered, or that too much information was remembered. Others expressed concerns for the agents forgetting details, hallucinating information, or having their personal information leaked elsewhere and no longer remaining private.

3.2 Heterogeneity of Memory and Temporal Dependency of Memory Needs

Our interview findings and coding of Reddit posts reveal that users do not consider all types of agent memory to be equal, and their needs for the system to remember certain information evolve over

time, depending on the task and the stage of the task that the user is working on with the agent.

For instance, participants described their memory needs as differing across domains, and this is one of the reasons why some participants create separate agents, personas, or custom instructions for different tasks. In general, participants wanted to organize and separate AI systems' memories clearly by task, project, and domain (e.g., separate memories related to writing from those related to health advice). Five of our interview participants (P1, P3, P4, P5, P6) even wanted memories from different projects, tasks, or domains to be disjoint from one another in the system, and when working on one task, they did not want the system to have access to memories from other unrelated tasks (e.g., they did not want the system to have access to memories about personal health advice when helping the user with academic writing).

> *"If I'm currently using it for, like, some project,* ***I would like it to recall the information for that project.***" — P5

Some participants did not want groups of memories to be completely disjoint though. For example, P4 thought that there could be benefit to the agent's memories partially overlapping across domains or tasks, and that the user could define how much overlap there is:

> *"Let's say,* ***I don't want my conversation about research to overlap with a side project conversation****, but I might still be using the same information. And so, just for convenience or visual representation, I open two separate chat windows for it. And in that case I would prefer it if the agent preserved those memories.* ***But I also think of it as something that the user should have more control over.***" — P4

A similar sentiment was shared in the Reddit discussions. From our analysis of both the interview and Reddit data, users tended to consider agents' memories as belonging to the following categories:

- **Factual memories:** Memories of general knowledge facts. These are often contained in the model's training data (parametric memory) or fetched through RAG [9].
- **User-related memories:** Memories related to the user, including their personal preferences (e.g., preferred conversational or interaction styles), personal details (e.g., their name, age, occupation, or other biographical information), personality, social, emotional, or other information.
- **Domain-related memories:** Memories related to a specific domain that the user wants the agent to specialize in. This could include any prompts or system instructions used to define the role or persona of the agent, or important domain-related terminology or concepts that the user feeds to the agent as context.
- **Project-related memories:** Memories of higher-level contextual details about a specific longer-term project—e.g., a work project, hobby project, or even a personal long-term goal (e.g., lose weight).
- **Task-related memories:** Memories of lower-level details about a specific task—e.g., a specific assignment, component to a project, or activity toward one's higher-level goal.

These categories can be thought of as existing on layers. For instance, *factual memories* would refer to general knowledge information that the agent would always have access to, that is not specific to any given task, and could be useful for any activity that the agent assists the user with. LLMs in their default state may only have access to *factual memories*, without the personalization or specialization that result from collecting memories from user interactions. One level below this would be where the personalization of the agent toward the user begins. *User-related memories* would refer to any general information that the agent knows about the user that could be applied in different ways to any task that the agent helps the user with. As some participants expressed though, not all user-related memories may be relevant to the task at hand (e.g., information about the user's family members may not be important for helping the user solve a programming task). Below this level would be where memories extend from user-level personalization to role specialization, and memories from this level downward influence the agent's specialization toward certain domains (*domain-related memories*), longer-term projects (*project-related memories*), and immediate tasks (*task-related memories*).

Participants expressed that their needs for the system to collect and make use of each of these types of memories evolves over time. For example, the user may need the system to recall a lot of memories about the current task with a great amount of detail and accuracy, while remembering only relevant details about other tasks, and/or recalling only higher-level (e.g., summarized) memories about other tasks.

> *"I can imagine, like, a scenario where, for example, I learned Haskell using GPT, and then I start a fresh chat to learn, let's say, Swift. And then GPT remembers what I know about functional programming and, let's say, lazy evaluation from the Haskell lectures. And so it might call back to that in this conversation, because Swift also has lazy evaluation, and that might be one way in which you can carry over long-term memories. Maybe if there's, like, a problem that I'm working on, and I switch to another problem, but GPT remembers that on the previous problem we encountered, like, a certain situation where I struggled, and I required extra assistance for it... so maybe it preemptively provides me with that information in [the] new chat, because it anticipates that I'll run into a similar problem.*" — P4

In addition, the user may also want the system to recall and make use of memories at higher-levels of the hierarchy (e.g., the user's personal details from the system's *user-related* memories), but only to the degree that is necessary or relevant for the specific task at hand.

> *"I mean,* ***[I do] not want it to remember my personal information a lot.*** *So, like, previously I said I want it to learn about my writing style, so that when I'm ready to revise my email, it is following my tone instead of using someone else's tone. [...] I don't want it to remember, like, my name. [...] So I don't want it to remember like, 'oh, I'm Tom's friend', 'I'm married', 'I'm a student', 'I'm, like, whose daughter', and 'I'm whose wife', like those kind of information, is not what I want it to remember.*

> *I just want it to remember like, 'oh, write it this way', and that's everything [it] should do."* — P6

3.3 Users' Practices of Organizing and Separating Memories

Participants explained that they attempt to separate memories by creating separate chat threads or sessions for different projects or tasks (P1, P4, P5, P6) or creating entirely new agents or personas (P2, P3, P4) by prompting the model with different base instructions (e.g., creating 'custom GPTs') to tailor the agent to specific roles, fine-tuning the model to give the system 'expertise' or specialization in a certain task or domain, changing the model in use (e.g., switching from GPT to Llama), or even logging in with different user accounts.

> *"For the HCI scholar [agent], like, that's just one for everything, like all of my writing academic writing that I have to do. And then for each project, I will create a custom one [agent], for example, like that usability one [project]. So if I need to evaluate the usability of one website, I'll create one [a custom agent]. And then I'll upload, like, to the knowledge [the agent's memory], like all of the screenshots for this one website. And if I'm working on like an app, now I'll create another one, and then like upload all of the screenshots for the app."* — P3

However, with many existing tools, when the user creates a new session or agent, they often need to manually select which information or memories they pass on to the new agent or new session. For example, the user might reuse the same base instructions or prompts defining the agent's role or higher-level task, but they might include different task or project-specific details for each instance of a task that they are working on.

> *"Yeah, I give it the same persona like, you know, 'you are an expert researcher, you have multiple years of experience, you're very familiar with, like, you know, some heuristics'. So that, like, persona blurb is the same. But then the blurb that I talked to them about the product. Like, this product, it contains XYZ... So I think, specific to this product, I'll upload different context for it to understand."* — P3

For some situations, this can be a cumbersome process. For example, some participants mentioned needing to repeatedly manually select relevant information to copy and paste across sessions or agents.

> *"Yeah, I would say, it's kind of difficult to do now. The only way I can keep it consistent is to copy and paste the same prompt."* — P2

> *"If it's not able to recall certain discussions that I had with [it], then I would try to like, feed again, [...] like, dump data, whatever I have.* ***Or maybe copy and paste the data from the previous chat [...] copy, paste the same conversation to it again.****"* — P5

Some participants wished that this process of selecting which memories to preserve or turn on/off would be easier.

> *"Yeah, it would be helpful if I could have like, you know... like, different settings, or like different fields that I can customize, right? Like, say, I want, like, a UX expert. Then I can have that as the standard, and then just swap out, like, other things."* — P3

Additionally, the Reddit community expressed many desires to tune memories and recommended features such as an on/off toggle, manually selecting, deleting, organizing and categorizing different memories. In actuality, users are *already* directly managing the system memory by using various on/off features, post-hoc editing, periodically pruning the memory, using custom instructions, explicitly telling the agent what to remember or forget, manually deleting from the chat history, repeating information that they want to be remembered, and even prompting the system for how it should remember going forward.

4 Design Opportunities

In contrast to the desires and practices expressed by our interview participants and those who posted in the Reddit discussions we analyzed, many existing systems attempt to separate or organize memories automatically via other means, such as through vector embeddings (where the system retrieves memories by linguistic or semantic similarities with the user's input(s); e.g., [13]), time-based storage and retrieval (where more recent memories are more likely to be retrieved that older memories; e.g., [13]), and by mimicking existing psychological models of human memory (e.g., by separating memories into *semantic*, *episodic*, and *procedural* memories [4]). While these are all promising approaches to implementing agent long-term memory, there is also opportunity to consider users' existing approaches to organizing information and system memories by project, task, and domain, and to consider how users' needs evolve as they transition from one activity to another. Therefore, users could also benefit from memory mechanisms that organize and control the level of access to or abstraction of memories based on the user's task-based needs.

Therefore, we recommend that agents, especially those that users use for multiple activities over a longer period of time, keep track of the different stages of the user's activities (task, project, domain, etc.) over time, and organize their memories based on these different stages of the user's activities. This could be done both with or without input from the user—e.g., the agent can recognize explicit or implicit cues from the user that they are working on a new task or domain, including from the user's own prompts (such as what task they explicitly say they are working on) or by observing how the user organizes their own chat threads by project or task.

There could also be benefit to allowing the user to directly manipulate and control the level of access to certain memories by task. In addition to existing approaches that allow users to directly manipulate agent memories (e.g., allowing users to move memories from one conversation to another or edit the level of abstraction of memories [15, 34]), allowing users to organize memories into hierarchical structures (e.g., akin to folders and subfolders in file systems, or global vs. local variables and private vs. public properties in object-oriented programming) where they can define the appropriate level of access to memories by task or activity could be a potentially promising approach that aligns with our participants' existing needs for agent memory.

Further, allowing users to organize memories into user-defined groups (such as, for example, *'personal memories'*, *'writing feedback memories'*, and *'health advice memories'*), and allowing users to toggle on/off and control the level of access to or abstraction of these user-defined groups of memories could be another way to give users more control, involvement, and agency over how the agent's memory works for them.

5 Conclusion and Future Work

This work provides preliminary understanding of users' long-term expectations and needs from AI systems' memory mechanisms, their practices of managing these systems' memories, and how their needs and practices evolve over time based on the stages of the activities they work on with the agents they use. To expand on these findings, we plan to run further studies (e.g., more in-depth interviews and longer-term studies with more participants) to better understand people's evolving needs over a longer period of time. We are also interested in understanding how certain design choices in AI systems (e.g., anthropomorphic design traits, direct manipulation vs. interface agents [28]) affect people's perceptions of and practices toward agent memory.

References

[1] [n. d.]. ChatGPT. https://openai.com/chatgpt/overview/

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

[3] 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.

[4] R. C. Atkinson and R. M. Shiffrin. 1968. Human Memory: A Proposed System and its Control Processes1. In *Psychology of Learning and Motivation*, Kenneth W. Spence and Janet Taylor Spence (Eds.). Vol. 2. Academic Press, 89–195. doi:10.1016/S0079-7421(08)60422-3

[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. doi:10.48550/arXiv.2210.08750 arXiv:2210.08750 [cs].

[6] 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.

[7] 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

[8] 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

[9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. doi:10.48550/arXiv.2312.10997 arXiv:2312.10997 [cs].

[10] 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

[11] 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.

[12] 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

[13] Yuki Hou, Haruki Tamoto, and Homei Miyashita. 2024. "My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents. 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:10.1145/3613905.3650839

[14] 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.

[15] 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 (UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–3. doi:10.1145/3586182.3615796

[16] 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

[17] 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].

[18] Taewoon Kim, Michael Cochez, Vincent Francois-Lavet, Mark Neerincx, and Piek Vossen. 2024. A Machine With Human-Like Memory Systems. doi:10.48550/arXiv.2204.01611 arXiv:2204.01611 [cs].

[19] 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.

[20] 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.

[21] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. MemGPT: Towards LLMs as Operating Systems. doi:10.48550/arXiv.2310.08560 arXiv:2310.08560 [cs].

[22] 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

[23] 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].

[24] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3586183.3606763

[25] 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

[26] 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.

[27] 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

[28] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.

[29] Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. 2024. Recursively Summarizing Enables Long-Term Dialogue Memory in Large Language Models. doi:10.48550/arXiv.2308.15022 arXiv:2308.15022 [cs].

[30] 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

[31] 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

[32] Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. doi:10.48550/arXiv.2107.07567 arXiv:2107.07567 [cs].

[33] 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. doi:10.48550/arXiv.2203.05797 arXiv:2203.05797 [cs].

[34] Ryan Yen and Jian Zhao. 2024. Memolet: Reifying the Reuse of User-AI Conversational Memories. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3654777.3676388

[35] 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.

[36] 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].

[37] 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. doi:10.48550/arXiv.2204.08128 arXiv:2204.08128 [cs].

[38] 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 (March 2024), 19724–19731. doi:10.1609/aaai.v38i17.29946 Number: 17.

A Semi-Structured Interview Guide

Interviews were semi-structured, and the exact questions asked occasionally deviated slightly from the following guide.

A.1 Introduction

(1) Share with me how you currently use personalized AI tools like ChatGPT, Claude, or Gemini.
   - **What AI tools** or chatbots do you use on a regular basis?
   - What do you use these tools for? (What types of **tasks/activities**?)

(2) Share with me the **role that memory plays** in your use of these systems.
   - What things does the system remember over time?
   - How does the system's memory affect the task(s) it does (or helps you do)?

A.2 Users' Mental Models of and Desires from Agent Memory

(3) Share with me your understanding of how the memory in [X system] works.
   - How does it **store** memory (how does it remember things)?
   - How does it choose **what to remember**?
   - How does it **recall** memories?
   - How does it **use** memories?
   - How does it **behave** based on its memories? How do you **expect it to behave** based on its memories?

(4) What kinds of behaviours from the system (e.g., indications, communications, interface elements, etc.) **help you better understand the system's memory operations**?
   - Is there anything that [X system] does that you think helps you better understand how the system's memory is working (e.g., what it is remembering, recalling, forgetting, etc.)?

A.3 Users' Practices with Agent Memory

(5) Is there anything that you do with the system specifically to **try to get it to remember certain things** (or forget certain things)?
   - Is there anything you do with the system to **try to *prevent* it from remembering certain things**?

(6) In what instances do you **want to be more aware** of what is in the system's memory?
   - In what instances do you want to know what the system is choosing to remember? (In what instances do you want to control this?)
   - In what instances do you want to know what and how the system is recalling information in its memory? (In what instances do you want to control this?)
   - In what instances do you want to know what the system is choosing to forget? (In what instances do you want to control this?)

(7) In what instances do you want to **manage the system's memory**?

A.4 Conclusion

(8) Is there anything else that you would like to share?

B Survey Questions for Interview Participants

(1) *[Optional; open response]* What is your **age**?
(2) *[Optional; multiple choice; select one]* Which **gender** do you identify as?
   - Female
   - Male
   - Non-binary
   - Prefer not to say
   - Other *[please describe]*
(3) *[Multiple choice; select one]* How often do you use **LLMs or personalized AI tools** (e.g., ChatGPT, Claude, Gemini, Pi, Meta AI, Microsoft Copilot chatbot, NotebookLM, Perplexity, AI-generated meeting summaries, AI-powered writing assistants, Copilot in Microsoft Office or GitHub)?
   - At least once per day
   - At least once per week
   - At least once per month
   - At least once per year
   - Have only used these tools a few times before
   - Have never used these tools before
(4) *[Open response]* Which of these tools do you use, and what do you use them for?
(5) How would you rate your **level of expertise** with using these tools?
   - *Likert scale from 1 to 5, 1 = "Beginner/novice", 5 = "Advanced/expert"*
(6) Please rate your level of **technical knowledge** of LLMs—e.g., your technical understanding of how LLMs work, your technical understanding of memory in LLMs and LLM-based tools and agents (including model training, fine-tuning, context windows, RAG, etc.).
   - *Likert scale from 1 to 5, 1 = "Beginner/novice", 5 = "Advanced/expert"*
(7) *[Optional; open response]* Please **describe / elaborate on your technical knowledge** of LLMs and LLM-based tools. What specific knowledge do you have about how these tools work?

C List of Subreddits where the Reddit Discussion Threads were Collected From

- r/Artificial
- r/ArtificialIntelligence
- r/CharacterAI
- r/ChatGPT
- r/ChatGPTCoding
- r/ChatGPTPro
- r/ChatGPTPromptGenius
- r/CyberSecurity
- r/EuroPrivacy
- r/ExplainLikeImFive
- r/Grok
- r/JanitorAI_Official
- r/LangChain
- r/LearnMachineLearning
- r/LLMDevs
- r/LocalLLaMA
- r/NoStupidQuestions
- r/OpenAI
- r/Privacy
- r/PromptEngineering
- r/Shortcuts
- r/Singularity
- r/Technology
- r/YouShouldKnow