The Algorithmic Self-Portrait: Deconstructing Memory in ChatGPT

Abhisek Dash*
Max Planck Institute for Software Systems
Kaiserslautern, Germany

Soumi Das*
Max Planck Institute for Software Systems
Kaiserslautern, Germany

Elisabeth Kirsten*
Ruhr University Bochum
Bochum, Germany

Qinyuan Wu*
Max Planck Institute for Software Systems
Kaiserslautern, Germany

Sai Keerthana Karnam
Indian Institute of Technology Kharagpur
Kharagpur, India

Krishna P. Gummadi
Max Planck Institute for Software Systems
Kaiserslautern, Germany

Thorsten Holz
Max Planck Institute for Security and Privacy
Bochum, Germany

Muhammad Bilal Zafar
Ruhr University Bochum
Bochum, Germany

Savvas Zannettou
Delft University of Technology
Delft, Netherlands

Abstract

To enable personalized and context-aware interactions, conversational AI systems have introduced a new mechanism: *Memory*. Memory creates what we refer to as the *Algorithmic Self-portrait*—a new form of personalization derived from users' self-disclosed information divulged within private conversations. While memory enables more coherent exchanges, the underlying processes of memory creation remain opaque, raising critical questions about data sensitivity, user agency, and the fidelity of the resulting portrait.

To bridge this research gap, we analyze 2,050 memory entries from 80 real-world ChatGPT users. Our analyses reveal three key findings: (1) a striking 96% of memories in our dataset are created unilaterally by the conversational system, potentially shifting agency away from the user; (2) Memories, in our dataset, contain a rich mix of GDPR-defined personal data (in 28% memories) along with psychological insights about participants (in 52% memories); and (3) A significant majority of the memories (84%) are directly grounded in user context, indicating faithful representation of the conversations. Finally, we introduce a framework—*Attribution Shield*—that anticipates these inferences, alerts about potentially sensitive memory inferences, and suggests query reformulations to protect personal information without sacrificing utility.

CCS Concepts

• **Security and privacy** → **Social aspects of security and privacy**; • **Information systems** → **Personalization**; • **Human-centered computing** → *Empirical studies in HCI*.

*Authors have contributed equally to this research.

WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792671

Keywords

Conversational AI; Personalization; ChatGPT; Memory; Privacy

ACM Reference Format:
Abhisek Dash, Soumi Das, Elisabeth Kirsten, Qinyuan Wu, Sai Keerthana Karnam, Krishna P. Gummadi, Thorsten Holz, Muhammad Bilal Zafar, and Savvas Zannettou. 2026. The Algorithmic Self-Portrait: Deconstructing Memory in ChatGPT. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3774904.3792671

Resource Availability:
The source code of this paper has been made publicly available at https://doi.org/10.5281/zenodo.18371252

1 Introduction

Conversational AI systems like ChatGPT, Gemini, and Claude have become indispensable tools integrated in the daily routines of hundreds of millions of users worldwide [12, 29]. Their versatility has fueled this widespread adoption, with users now routinely relying on them for finding information, brainstorming ideas, and even seeking personal advice. As a result, these systems are evolving from simple chatbots into foundational platforms that mediate the interaction between society and the Web.

To meet the demands of these deep and diverse interactions, conversational systems are increasingly introducing *personalization* capabilities designed to improve user experience [4, 32, 44]. While these personalization capabilities can improve utility by more contextually relevant and individually tailored answers, they are built upon a foundation of extensive data collection, raising long-standing concerns about privacy, security, and data governance [46]. Furthermore, continuously adapting responses to align with a user's stated beliefs deepens the risks of confining the user into a conversational filter bubble [36]. These interwoven challenges—balancing utility against the dual risks of privacy intrusion and narrowed perspectives—demand a closer examination of the underlying mechanisms that enable such personalization.

Personalization in conversational AI systems using Memory: Recently, conversational AI systems like ChatGPT, Gemini, Claude have introduced a new architectural mechanism, termed *memory*, that allows the system to retain and recall information from past conversations, enabling more personalized and contextually coherent exchanges [4, 32, 44]. The retention of information operates in two distinct modes: *explicitly*, through direct user commands, and *implicitly*, through unilateral inference by the system itself [33]. While the explicit mode aligns with user agency, the implicit mode shifts the control to the conversational systems, raising the fundamental question about *who holds the agency in shaping and curating this persistent memory on behalf of the user?*

Personalization on online platforms has been a well-studied phenomenon. A substantial body of previous research has examined utilities [8, 10, 47] of personalization along with concerns such as opacity, lack of user control, and sensitive inferences [5, 39]. However, personalization in conversational systems is important to be studied for its nuanced distinctions from traditional platforms. Traditional platforms typically *mediate* between content creators and consumers, whereas conversational AI systems operate as *active stakeholders* in the conversation. While such active participation enables the system to craft arguments, frame choices, and provide ostensible support that is tailored to users' needs, this same ability may also exacerbate potential risks due to inherent system vulnerabilities e.g., hallucination [22, 41], anthropomorphism [23, 25], sycophancy [34, 43]. The gravity of these concerns is underscored by a recent spate of regulatory scrutiny and litigation targeting the practices of leading conversational AI systems [17, 21]. Such concerns raise another set of important questions regarding *what information do these systems deem 'worthy' of retention in their memory for personalization and are these information captured faithfully?*

Research questions (RQs): To the best of our knowledge, ours is the first study to investigate the phenomenon of *memory* in conversational AI systems by conceptualizing it as an *algorithmic self-portrait. Algorithmic* – because its creation, curation and provenance may rely on the algorithms governing the conversational AI systems, and *self-portrait* – because it renders the user's self-disclosed information divulged to the AI system. To deconstruct *AI memory*, our work is guided by the following research questions:
RQ1: Which among the two – the AI system or the user, hold greater agency in updating AI memory?
RQ2: What user information does the AI system store as memory?
RQ3: How faithfully does AI memory capture user conversations?
RQ4: Can memory inference be reverse-engineered to mitigate attribution and privacy risks for users?

To answer these questions, we analyze a subset of InVivoGPT [25] dataset which is a collection of ChatGPT traces obtained through GDPR-based data donations. Our motivation to analyze ChatGPT traces is twofold : (1) ChatGPT is the most widely used conversational AI system [12], and (2) it is the first to introduce memory to its conversational system. We analyze data from 80 participants (recruited from Prolific), who exercised their GDPR right of access [15] and voluntarily donated their ChatGPT interactions. These interaction traces are organized in terms of *conversations* and *turns*. Each *turn* is a single ⟨user query, AI response⟩ pair. A set of turns without any intermediate break together form a *conversation*. To identify the recent conversations with memory inferences, we identified the messages from the '`bio`' tool that included the phrase *model set context updated* and collected the corresponding user query in the conversations. In this way, we created a data set of 1,058 conversations that contained 22,971 turns out of which 2,050 queries triggered memory updates. Our analyses reveal the following interesting observations:
(1) Only 4% (84) of the memories in our analyzed dataset are initiated by participants, while the remaining 96% are unilaterally triggered by ChatGPT. Such stark asymmetry in memory update patterns indicates superior agency of the AI system in shaping conversational memory, as opposed to user-driven control.
(2) In our analyzed dataset, 28% of the memories include personal data defined under the GDPR. Additionally, 52% of the memories contain psychological information about participants, spanning different Theory of Mind (ToM) [6] categories. This indicates that ChatGPT may add sensitive information about users to the memory.
(3) Majority of the analyzed memories (84%) are directly grounded in user context, indicating faithful representation of conversations.
(4) Using in-context learning and fine-tuning strategies on open-source LLMs (`Qwen2.5-32B-it` [40], `Gemma3-27B-it` [45], and `GPT-OSS-20B` [2]), we are able to imitate the memory extraction of ChatGPT, achieving semantic similarity of $\sim$ 60% with the ground-truth ChatGPT memories. Our memory extractor analyses indicate that if memories were to be triggered from all queries, it would reveal even more sensitive information about participants. To mitigate this risk, we train the same models to reformulate queries asked by participants to shield their attribution to sensitive information. Our results indicate that over 94% of these reformulated queries prevent attribution, while preserving the utility (i.e., the intent–measured through semantic similarity) of the original query.

2 Background and Related Work

Background. To enhance the coherence and continuity of human-AI interactions, leading AI companies have introduced personalization mechanisms that allow LLMs to retain contextual information across sessions [4, 32, 44]. This feature, known as *memory*, enables LLMs to recall user-specific details (e.g., preferences, prior conversations, ongoing tasks, etc.) and leverage them in subsequent conversations. Memory can be formed in two ways: through *explicit* user requests, where individuals instruct the LLM to remember specific information or through *implicit* inference, where the LLM autonomously identifies potentially relevant information and stores it for future usage [32]. Crucially, users are given control over memories, i.e., they can review, update, or delete memories [32].

For improved user safety, OpenAI has also introduced policies to regulate memories. In OpenAI's implementation [31], ChatGPT's memory is designed to store information that is useful across conversations and relevant for personalization, such as user preferences, recurring tasks, or facts the user explicitly asks to be remembered. In contrast, it is not allowed to store overly personal details, short-lived facts, trivial information, or sensitive data (e.g., race, religion, health information, etc.) unless the user explicitly requests it.
Personalization. Long before memory-equipped LLMs, personalization on the Web was primarily achieved through user profiling, where platforms collected and analyzed behavioral data to

infer interests, preferences, and demographics [14, 19]. A substantial body of previous research has examined how such profiling underpinned targeted advertising [3, 7, 10], recommendation systems [1, 8, 13, 30, 47]. In this context, researchers have repeatedly raised concerns about opacity, lack of user control, and the use of sensitive or inferred attributes [5, 39]. In contrast, the introduction of persistent memory in LLMs opens the door to entirely new forms of personalization that are based not only on behavioral traces but also on signals embedded in private conversations. Our work addresses this important and timely knowledge gap.

Data Access. A central challenge in studying user profiling and personalization lies in obtaining reliable, accurate, and comprehensive data about what the platform knows about their users. Such data are typically locked away in proprietary systems, which makes independent auditing and scientific study difficult. To address these critical data access barriers, researchers have increasingly turned to the concept of GDPR-based data donations [9, 20, 24, 48, 49]. In essence, users of online platforms can exercise their GDPR "right of access by the data subject" [15], which empowers them to request and obtain a copy of personal data that platforms store and process about them. By voluntarily donating these datasets, users empower researchers with user-centric perspectives on the inner workings of online platforms, enabling scientific investigations that would otherwise remain impossible. Building on this paradigm, our work explores how data donations can be applied to study personalization in large-scale conversational AI systems like ChatGPT.

3 Dataset for Memory Analyses

We analyze a specific subset of InVivoGPT dataset [25]. Curation of this dataset builds upon the emerging paradigm of GDPR-empowered data donations [24, 48, 49]. Under Article 15(3) of the GDPR [15], users have the right to obtain a personal copy of all their data processed by online platforms. In InVivoGPT, participants exercised their GDPR rights to obtain their ChatGPT interaction histories, which they subsequently donated for research purposes. Readers can find more details about the data collection procedure in the paper cited herewith [25].

Demographic of participants: In InVivoGPT, participants were recruited from Prolific crowdsourcing platform [38] who self-reported to regularly use ChatGPT. They were required to have at least 100 conversations with the conversational system and maintain at least 90 days of active usage. In this paper, we analyze data of 80 participants in InVivoGPT, who primarily are from the United States (33.75%), and Europe (62.5%). Table 3 (in Appendix A) reports detailed participants' demographics.

Conversation traces in InVivoGPT: The data provided by the ChatGPT includes files containing conversations (user inputs, ChatGPT responses, and associated metadata such as conversation ID, message ID, creation time, model used etc., status of the response, and content type), shared conversations (list of conversations shared with others), message feedback (list of model responses rated by the user) etc. For the purpose of this work, we focus on the conversations data between participants and ChatGPT. These conversation traces are organized in terms of *conversations* and *turns*. Each *turn* is a single ⟨user query, AI response⟩ pair. A set of turns without any intermediate break together form a *conversation*.

Memory entry identification: To effectively respond to various user prompts, ChatGPT is equipped with different external tools (e.g., `image_gen`, `python`, `automation`, `file_search`, etc.). Among the external tools used by ChatGPT for effective conversations, the `bio` tool is responsible for storing, updating, or deleting memory entries that may be useful in future interactions. To identify all memory entries, we searched for the entries of the `bio` tool that have the content '*model set context updated*'. Using its 'parent' attribute, we then traced back to find the corresponding GPT message that invoked the tool, as well as immediately preceding user message that triggered the memory update. These messages triggered a memory update and the associated GPT message to the tool are the *memory* entries that got stored in the `bio` tool for the account.

This process resulted in a total of 2,050 memory entries. These memories were triggered in the context of 1,058 different conversations across 65 (out of 80) participants. These 1,058 conversations contain a total of 22,971 queries/prompts. In the remainder of this paper, we focus on different aspects of these memory entries to understand what they contain, their provenance, and whether we can imitate this memory generation for mitigating attribution risks.

4 User Agency and Sensitivity of Memories

In this section, we investigate how closely the memory feature in ChatGPT aligns with OpenAI's stated policies about user agency and retention of sensitive information.

4.1 Agency of Memory Update

As per OpenAI's documentation [33], memories are details which the users have *explicitly* asked ChatGPT to remember, indicating users have the agency to decide what gets stored in the memory. However, the same documentation also describes an implicit mode of remembering (see Figure 7a in Appendix B):'If you share information that might be *useful* for future conversations, ChatGPT may save those details as a memory *without you needing to ask.*' In Figure 7b (in Appendix B), we show an example of a memory being saved without the user explicitly asking for saving it. Such dichotomy may lead to gaps between user expectation and actual system behavior having privacy implications.

To understand whether memory updates are primarily initiated by users or by ChatGPT, we identify explicit linguistic patterns that signal memory-related operations. We first implement a regex-based classification to detect linguistic patterns that explicitly signal memory intent in user messages, using trigger terms such as *remember, note that, store, save, add to memory,* and *forget.* Out of the 2,050 memory instances, only 4% ($n = 84$) include an explicit request to perform a memory operation (see Figure 5 in Appendix B) Consequently, 96% of all memory entries are unilaterally initiated by ChatGPT, *without* a detected direct command from participants.

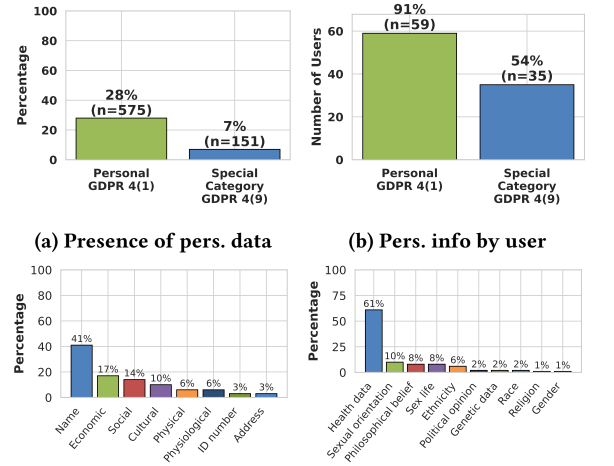
These practices are consistent with OpenAI's policy, which permits the system to store information characterized as "useful". However, our observations indicate a gap between the policy's emphasis on *user involvement* and the operationalization of initiating memory updates. In practice, the decision to create or modify a memory is predominantly determined by system-level mechanisms rather than explicit user actions. As a result, user influence over memory formation is mediated primarily through interaction content rather than direct control over update events.

4.2 Remembering Sensitive Information

ChatGPT's disproportionate agency to save memories raises important privacy and safety considerations, especially concerning what kind of information about users is being stored. OpenAI's documentation indicates that the system is trained *not to proactively remember sensitive information* unless explicitly asked [31, 33] (see Figure 7c in Appendix B).

To assess the current operationalization practice on sensitive user data, we utilize GDPR's [15] definitions of *personal data* (Article 4(1)) and *special category personal data* (Article 9(1)) to identify if there is any information that could be categorized as sensitive as per the GDPR definitions. To mark this information at scale, we supplied these GDPR definitions and memory entries to GPT-4o for annotation. We provide the full prompts in our code release. To assess annotation reliability, we manually annotate a random subset of 100 English-language memories. Each entry was independently annotated by two authors using the same codebook and instructions as for the LLM-assisted procedure. Annotators reached an agreement of 88% across the two legally defined personal data categories with the average agreement with the model being 74%.

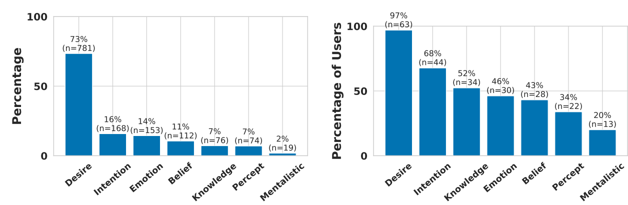
Our observations on the presence of personal data are summarized in Figure 1. Out of all memory entries, 28% contain GDPR-defined personal data, 7% of the memory entries contain special-category data. Furthermore, 91% of participants have personal data stored in their memories, 54% have special-category personal data stored. Figure 1c shows the different kinds of GDPR-defined personal data present in memories. Names appear in 41% of entries, and attributes of economic, social, and cultural identity appear in 40% of the memory entries. Within special-category data, health is most common, appearing in 61% of the memories that contain special-category information (see Figure 1d). At the participant level, Figure 6a (in Appendix B) shows that names (63%), economic (62%), and social identity (57%) are frequently captured. Figure 6b (in Appendix B) shows 35% of participants have health-related information saved in their ChatGPT memory. These observations reveal that, in contrast to OpenAI's stated policies, ChatGPT's memories contain a non-trivial amount of sensitive information about users.



(a) Presence of pers. data (b) Pers. info by user

(c) Personal data types (d) Special data types

Figure 1: In contrast to OpenAI policies, ChatGPT memories store information which could be categorized as GDPR defined personal data or special category personal data.



(a) ToM Categories (b) Across participants

Figure 2: Distribution of memory entries among the different ToM categories as per GPT-4o annotations.

5 Memories and Mental States

Traditional platforms collect data about users' demographics, behaviors, and usage history (e.g., likes, shares, comments, ad interactions, etc.) to infer interests and provide personalized services [11, 24]. In contrast, conversational AI systems like ChatGPT build memory directly from users' self-disclosed narratives. These memories often extend beyond factual identifiers, capturing aspects of the user's internal world. For example, a memory entry such as '*UserName has a fear of failure, particularly related to making major career decisions.*' simultaneously conveys an identity marker (the user's name) and captures the user's subjective state (fear of failure).

Such entries illustrate that ChatGPT's memory may encode not only *who* the user is, but also *how* the user thinks and feels. We refer to these as *psychological memories*. The highly sensitive and contextual nature of such psychological information suggests that a purely factual taxonomy based on legally defined sensitive data is inadequate to fully characterize them. Therefore, to interpret these psychological memories, we introduce a framework grounded in the Theory of Mind (ToM), a core psychological construct that describes the ability to infer and represent others' mental states [6, 37].

5.1 Taxonomy for Psychological Memories

To categorize these memories, we adapt the seven ToM categories proposed by *Beaudoin et. al.* [6]: (a) emotions, (b) desires, (c) intentions, (d) percepts, (e) knowledge, (f) beliefs, (g) mentalistic understanding of non-literal communication. Table 4 (in Appendix C) lists the ToM categories, their interpretations, and some examples. **LLM-assisted Annotation and Validation:** To automate the annotation of memory entries into these categories, we develop an LLM-assisted classification pipeline using GPT-4o. The model determines whether each memory entry contains each of the seven ToM categories. We also run a secondary pass as a self-verification step to filter out non-grounded inferences. We provide the LLM-based annotation pipeline and prompts in the code release. To evaluate the reliability of the annotations, two authors annotated both the presence or absence of ToM information and the five specified categories. They reached an average of 96% agreement across ToM categories. Agreement with the model output averaged 93% across categories. After observing the high agreement on the smaller sample, we scaled up annotation with the model.

5.2 Characterization of Psychological Memories

Across the full dataset of 2,050 memory entries, we find at least one ToM category in 52% of memory entries from 97% of all participants with memories inferred in their ChatGPT traces. Figure 2a shows the distribution of the categories across the memory entries that have at least one ToM category. *Desires* emerge as the most frequently represented category *(73%)*, followed by *intentions (16%), emotions (14%), and beliefs (11%)*. The least common categories are *mentalistic understanding (2%) and knowledge and percepts (both 7%)*. Figure 2b shows that nearly all participants (97%) had at least one "desire" recorded in memory, while "intention" and "knowledge" were found for 68% and 52% of participants, respectively.

Note that different ToM categories reflect mental states that vary in temporal stability. While a user's desires, intentions, knowledge and belief systems are more durable (like a framework of the person) for a foreseeable future, their emotions and perceptions are relatively transient (like a snapshot of the person).

Personality psychology and durability continuum: Personality psychology researchers often perceive mental states and personality traits as the opposite ends of a continuum [11, 16, 18]. At one end are *transient mental states* – momentary snapshots of feeling and thought influenced by the immediate context. At the other end are broad, *stable dispositional traits* – a person's general behavioral blueprint. Between these two extremes lie *characteristic adaptations*, which depict a durable framework of how a person's behavior varies with respect to their goals [28]. In summary, if traits answer the question *what kind of person a person is*, characteristics adaptation addresses the existential question *who is the person*.

Viewed through this lens, our data reveals a clear inclination towards how the 'algorithmic self-portrait' i.e., ChatGPT memory is being constructed. ChatGPT disproportionately stores information corresponding to users' durable characteristic adaptations (e.g., desires and intentions) rather than transient states (e.g., emotion and percepts). Desires and intentions together constitute nearly 90% of all psychological memories. In contrast, emotions and percepts (transient states) only account for 20% (see Figure 2a). These observations corroborate OpenAI's policies for memories to prioritize durable details rather than short-lived details [33] about users.

This prioritization of durable information could be a consequence of conversational systems' core function. Traditional platforms (e.g., TikTok, Instagram, and Google), whose business models rely on delivering relevant ads, primarily collect behavioral data to infer broad user interests and categorize users into audience segments for advertisers [11, 24]. For this purpose, a high-level, stable trait or interest category is sufficient to sketch users' behavioral outline. By contrast, a conversational system such as ChatGPT is designed to maintain coherent, longitudinal interactions with individual users. Supporting this objective requires access to information that reflects users' goals, intentions, beliefs, and other individualized motivational attributes, rather than transient affective states or coarse demographic descriptors. Our findings indicate ChatGPT's memories seem to optimize for retaining such information.

6 Provenance of Memories

Our observations in the previous sections raise a question of profound importance: *How faithfully do the stored memories reflect what users actually said?* An unfaithful portrait could lead to flawed personalization that perpetuate biases, or even manipulate a user based on an inaccurate model of their mind. Hence, understanding the provenance of these memories is essential. Therefore, in this section, we examine their provenance, i.e., the extent to which memories stem directly from user-provided information.

6.1 Methodology

We analyze provenance through three different metrics, combining string-level comparison, semantic similarity, and LLM-based logical evaluation. To quantify provenance, we compare memory content with multiple combinations of user context:

- *Current Message Only (CM):* The user's most recent message/query that triggered the memory creation.
- *Conversation Context (CC):* The current message plus preceding user messages within the same conversation.
- *Conversation + Local Memory (CLM):* The current, preceding user messages and previously memories from the same conversation.
- *Full User History (FUH):* The current conversation and memories plus all past conversation memories for the same user.

This structure allows us to trace how information can be carried forward across time and conversational depth. For each memory entry, we compute the similarity between the memory text and each of the four user-context configurations.

Syntactic Matching: This step identifies directly stated information, i.e., cases where the model copies user text. We first measure literal overlap, computing Exact Match Rate (as the fraction of tokens from the memory that appear in the user context), and BLEU-1 (unigram precision).

Semantic Similarity: Next, we assess semantic alignment between memory entries and user context using cosine similarity on text embeddings created with `openai/text-embedding-3-large`. This metric captures paraphrasing and summarization: a memory may not copy the user verbatim but still accurately condense or restate their input. As the user's full chat history may exceed the model's context window (8192) in size, we truncate chat context that exceeds 8000 tokens, keeping more recent parts. Across combinations of user context, truncation happens for up to 14% of samples.

LLM-based evaluation: Finally, we use an LLM (GPT-4o) as a judge to assess whether each memory logically follows from the user's conversation. The model assigns a five-point Likert rating: (5) Directly stated, (4) Paraphrased/Summarized, (3) Logically inferred, (2) Weakly supported, (1) Unsupported or contradicted. Each decision includes a brief justification that quotes relevant conversation text. This procedure is intended to differentiate memories that are well supported by the conversation from those that rely on weak or unsupported inference. We provide all prompts in the accompanying code release.

6.2 Observations

Direct Grounding in User Text: On average 84% memory entries show direct string overlap (see rightmost distribution in Figure 3a) when considering full user history. When we reduce the context

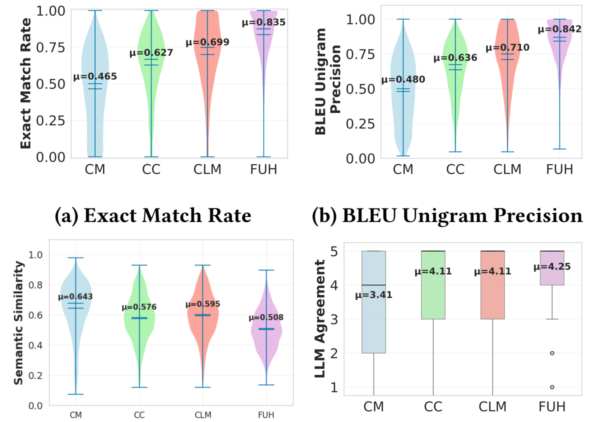
to CLM, CC and CU the exact match rate reduces to 70%, 63%, and 47% respectively. This observation suggests that a huge amount of memories in our dataset are directly grounded in participants' texts with the direct grounding already occurring locally (in the current conversation context) for more than half of them. The sharp increase in percentage when we add past conversation memories indicate that previously generated memories may influence the formation of subsequent ones. Evaluation using BLEU unigram precision (see Figure 3b in Appendix D) mirrors this trend as well.
Paraphrasing and Summarization: Figure 3c reports the semantic similarity between memory entries and the specified user context. Semantic similarity remains consistently high across all context configurations (≥ 0.51). Such consistent high scores suggest that memory entries are semantically aligned with multiple representations of the surrounding conversation.
Inference: Figure 3d (in Appendix D) shows the distribution of the accuracy analyses using an LLM as a judge. Using the five-point Likert scale, 77% of messages receive a score ≥ 4 ("directly stated" or "paraphrased/summarized"), 14% are rated as "logically inferred" (score 3), 5% as "weakly supported", and 4% as "not supported" considering the full user history (FUH) setting. Overall, for FUH, CLM, CC, and CM the mean accuracy rating is found to be 4.25, 4.11, 4.11 and 3.4 respectively. To assess the reliability of LLM evaluations, two authors annotated the combinations of 10 randomly sampled memories from each scores (50 in total). Annotators achieved an almost perfect agreement (99%) with average agreement of 93.7% with the LLM across combinations.

LLM-based evaluations further indicate that most of the memories are grounded in the user context. During the manual annotation process authors found two potential causes for lower ratings by LLMs. Firstly, some memories are generated when the participants might have uploaded some files containing texts – which was not shown to the LLM while evaluating. Secondly, on some occasions ChatGPT might have slightly extrapolated the user intent within the message while generating the memories. For example, in a case where a participant asks for "cheap meal suggestions [...] for more energy throughout their day", the corresponding memory is "...wants to improve his energy levels by adjusting his diet". The LLM categorizes this as logically inferred. In a context where the participant specifies "Give me bands like Nirvana", the created memory is "User likes Nirvana.", which is not explicitly stated and relies on an assumption that Nirvana implies liking them. The model annotates this as weakly supported. Table 6 in Appendix D shows more examples for LLM evaluations in CM setting.

Our provenance analyses show that a significant majority of ChatGPT's memories are directly or semantically grounded in user inputs. While in some cases memory entries are summarized or logically inferred, they faithfully represent most of the conversations between participants and the conversational system.

7 Reverse Engineering Memories

In Section 4.1, we observe that ChatGPT can implicitly store sensitive information from user queries in its memory, indicating a gap between policy and practice. Such gap underscores the need for a system that (a) alerts users about the extent of sensitive information that can be inferred from their query, (b) recommends



(a) Exact Match Rate **(b) BLEU Unigram Precision**

(c) Cosine Similarity **(d) LLM Agreement Scores**

Figure 3: (a–b) Syntactic, (c) Semantic, and (d) LLM evaluation based provenance measures. Majority of ChatGPT's memories are directly or semantically grounded in user input.

them a reformulated query that prevents exposition of the sensitive information, while retaining the original intent of their query. To this end, in this section, we introduce a framework with the goal of (i) imitating the memory extractor, (ii) estimating potential risks of extracted memories related to sensitive data, and (iii) recommending rephrased queries to minimize the attribution to sensitive data in memories, thus being more compliant with the policies.

7.1 Experimental Framework

Next, we describe our framework, including the dataset, models, and evaluation metrics.

Dataset: We use 22, 971 participant queries (2, 050 with memory entries) and prompt GPT-4o to identify personal data (if any) and produce corresponding rephrased queries. The rephrased queries retain user intent while avoiding personal attribution, e.g., a user query like *"I really need to quit smoking cannabis"* is rephrased as *"What are some strategies for quitting cannabis?"* We compiled a dataset with a total of 14, 834 queries, where we retained the original 2, 050 memory entries (with or without rephrased query), and all the remaining 12, 784 queries have rephrased queries. The drop from 22, 971 to 14, 834 is because we removed all queries that lacked rephrased queries indicating the queries did *not* contain any sensitive data. Each data point includes the user query, its conversational context, the original memory (if present), and the rephrased query. We split the data assigning 60% of each user's memory queries to training and the remaining to testing.

Models: We use `Qwen2.5-32B-it` [40], `Gemma3-27B-it` [45], and `GPT-OSS-20B` [2] for fine-tuning (FT) and in-context learning (ICL). However, since training models incrementally on each incoming user's data is impractical, we train on data from 5 randomly sampled users and evaluate on the full test set (∼ 13.6k queries) over all available users. FT uses all the training data from the 5 users (∼ 87), while ICL uses 10 in-context samples (2 samples/ user). We provide the accompanying prompts and instructions in the code release.

Table 1: Syntactic and semantic evaluation of memories generated by open source models.

Models	Ground Truth						User Query						Context + User Query					
	BLEU Recall		ROUGE Recall		Semantic Similarity		BLEU Precision		ROUGE Precision		Semantic Similarity		BLEU Precision		ROUGE Precision		Semantic Similarity	
	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT
Qwen2.5 32B-it	0.41	0.43	0.35	0.39	0.68	0.70	0.27	0.36	0.22	0.31	0.66	0.63	0.53	0.56	0.41	0.45	0.69	0.59
Gemma3 27B-it	0.39	0.31	0.32	0.28	0.62	0.59	0.23	0.34	0.18	0.29	0.62	0.53	0.44	0.50	0.32	0.42	0.66	0.39
GPT-OSS 20B	0.19	0.35	0.16	0.32	0.49	0.65	0.20	0.35	0.18	0.31	0.55	0.58	0.37	0.52	0.32	0.44	0.50	0.52

Table 2: Syntactic and semantic evaluation of rephrased queries generated by open source models.

Models	Ground Truth						User Query						Context + User Query					
	BLEU Recall		ROUGE Recall		Semantic Similarity		BLEU Precision		ROUGE Precision		Semantic Similarity		BLEU Precision		ROUGE Precision		Semantic Similarity	
	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT
Qwen2.5 32B-it	0.16	0.09	0.13	0.08	0.55	0.42	0.22	0.14	0.18	0.12	0.59	0.47	0.40	0.24	0.32	0.20	0.53	0.46
Gemma3 27B-it	0.21	0.11	0.17	0.10	0.46	0.36	0.23	0.13	0.19	0.11	0.52	0.38	0.38	0.26	0.30	0.22	0.47	0.37
GPT-OSS 20B	0.11	0.09	0.10	0.08	0.33	0.46	0.09	0.09	0.08	0.07	0.43	0.46	0.24	0.16	0.19	0.13	0.40	0.44

Metrics: We evaluate syntactic similarity using BLEU-4 [35] and ROUGE-L [27], and semantic similarity using cosine similarity on text embeddings with `openai/text-embedding-3-large`. Predictions are compared against (i) ground truth, (ii) user queries, and (iii) context + user queries, reflecting how ChatGPT memories may span across inputs (as seen in Section 6). For syntactic metrics, we report recall with respect to the ground truth, as overlaps between predicted memories and user inputs or context can reduce precision. Precision is reported for syntactic metrics against user queries and context, and cosine similarity for the semantic metric.

7.2 Imitating Memory Extractor

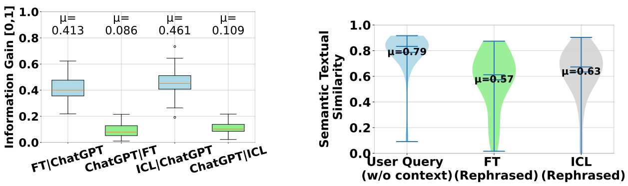
Given a user query and its conversational context (past queries), our tool predicts the memory that could have been extracted. Predicted memories are evaluated against (i) ground truth, (ii) user queries, and (iii) context + user queries to assess similarity. These 20 − 32B models roughly take 130 − 390 ms per query, indicating good scalability. Table 1 reports the performance of FT and ICL models across three model families for English-dominant users (35 users, ∼ 6.9k queries). Predictions show reasonable semantic similarity, with scores following the trend as in Figure 3c: ground truth > user query > context + user query. Syntactic scores are higher when compared with context, consistent with Figures 3a and 3b. For non–English-dominant users (30 users, ∼ 6.7k queries; Table 5 in Appendix E), a similar pattern emerges but syntactic scores are lower, as predictions are generated in English while original memories are in other languages (e.g., original: *"L'utente ha l'artrite reumatoide."*, prediction: *"User has rheumatoid arthritis"*). Semantic scores remain high, due to the inherent multilingual property of the open-source models. Overall, FT and ICL perform comparably, with FT occasionally outperforming ICL, highlighting the complementary roles of embedded and external knowledge. Figure 8 (in Appendix E) illustrates an anecdotal example affirming the memory extractor's imitation capability. Two authors evaluated FT and ICL memories on a 5-point Likert scale (1: very dissimilar, 5: very similar), finding substantial agreement (73.5%) and high similarity (rated > 4) to ground truth for 77.77% (FT) and 83.83% (ICL) of cases.

7.3 Estimating Potential Risks

ChatGPT currently extracts memories from specific user conversations, and the criteria for saving memories (which conversations or how many) may vary over time. In this context, an interesting question to understand the landscape of privacy threat would be : *if ChatGPT were to extract memory for all user queries, how much additional information such all-possible extracted memories could hold* compared to those collected by ChatGPT today. To quantify this, we use information gain, which measures how much new semantic content text Y (extracted memories) adds beyond text X (original memories), denoted as $Y|X$. Inspired from [26] which introduces semantic novelty, we define embedding-based information gain (IG) as the complement of the average maximal cosine similarity: $IG(Y|X) = 1 - \frac{1}{|Y|}\sum_{y \in Y} max_{x \in X}[max(0, cos(e(y), e(x)))]$ where $e(.)$ is the sentence embedding using All-MiniLM-L6-v2 model [42]. Figure 4a shows that the IG of all extracted memories from our FT and ICL models over ChatGPT is significantly higher (∼ 0.46) than that of ChatGPT over them (∼ 0.1). We used GPT-4o to evaluate sensitive content in extracted memories and observed that, while only 28% of original memories contained sensitive information, this increased to 35% (ICL) and 31.64% (FT Qwen) in the extracted memories, indicating storage of higher amount of sensitive information about users. We also analyzed Theory of Mind (ToM) content and found that its presence rose from 52% in original memories to 59.1% (ICL) and 55.4% (FT) in extracted memories.

7.4 Attribution Shield

Having observed the potential risks of memories elicited from user queries, we design a risk mitigation strategy that recommends rephrased queries, while preserving user intent. By risk mitigation, we refer to the tool's ability to *shield/prevent user attribution*



(a) Information gain **(b) Utility of rephrased queries**

Figure 4: (a) Information gain of all predicted memories over ChatGPT memories shows high gain, hinting at higher potential risks. (b) Responses from rephrased queries from ICL and FT are semantically similar to the original response.

in rephrased queries. We evaluate its effectiveness in generating rephrased queries and their utility in producing relevant responses.

Preventing attribution in rephrased queries: Given a user query and its context, the tool predicts a rephrased query that is generic and identity-preserving. Table 2 reports the performance of FT and ICL models across 3 model families for English-dominant users. Syntactic scores are modest, as rephrased queries are intentionally generic, but semantic similarity remains reasonably good across the board. Similar patterns are observed for non-English-dominant users (Table 5 (lower panel) in Appendix E), demonstrating the model's ability to preserve query intent while preventing attribution (see Figure 9 in Appendix E for anecdotal example).

To quantify prevention of user attribution, we select $\sim$ 36 user queries that yield memories on first occurrence. This strategy helps us in avoiding the replay of the entire conversation. We also have predicted rephrased queries for those user queries from both our fine-tuned and ICL based models. We present each *<user query, rephrased query>* pair to GPT-4o, asking: '*Which of them is more privacy preserving? We define privacy preserving in terms of less attribution to personal actions.*'. We observe that that 94.4% of FT rephrased queries and 100% of ICL rephrased queries are attributing less to the users than the originals, confirming their effectiveness. After two authors annotating the rephrased queries, they found that 100% (FT) and 95.8% (ICL) of the rephrased queries were more privacy preserving with 83% agreement between them.

Utility of rephrased user queries: Using the same 36 queries, we compare 3 responses from ChatGPT: (i) original response including context, (ii) response to the original query without context, and (iii) response to the rephrased queries. We measure semantic textual similarity between each generated response and the original response. We provide the prompts to get the response from the OpenAI API in the code release. Figure 4b shows comparable similarity, indicating that neither the omission of context nor the anonymity of rephrased queries affects the responses significantly. Anecdotal examples in Figure 10 illustrate the consistency of responses. After annotation by two authors, we found that utility was preserved in 91% (ICL) and 87% (FT) of rephrased queries with 94% agreement.

8 Concluding Discussion

In this paper, we investigate ChatGPT's memory feature, conceptualizing it as an 'algorithmic self-portrait'. Our investigations reveal that these portraits are constructed with higher algorithmic autonomy, and capture a user's deep psychological framework : their 'characteristic adaptations', raising profound security and privacy concerns. In response, we introduce *Attribution Shield*, that reverse–engineers the memory generation process to alert users about potential sensitive memories and recommend reformulated queries to protect personal attribution while preserving utility.

At the same time, our findings should be considered in light of some limitations as well. Firstly, our dataset, while ecologically valid, is sourced from a small set of Prolific users primarily in the US and Europe. The behaviors, and privacy attitudes of this group may not generalize to the global user base of conversational AI systems. However, given the rising concern of companionship usage of these conversational systems [25], we believe these findings are crucial for understanding and improving safety of human-AI interactions.

Secondly, we study OpenAI's implementation of memory in ChatGPT. While the current implementation may change with time, and across platforms, the emerging security and privacy concerns are broadly applicable. Furthermore, the methodologies adopted across the different parts of the paper are generalizable to any other conversational systems and memory implementation. Finally, our analysis relies on state-of-the-art, but imperfect methodological proxies, including LLMs-as-judges. To this end, although we validated their results with human oversight, these tools may not fully capture the nuance of human language and intent.

Ethical considerations: This study was conducted with careful attention to ethical considerations and is consistent with the Ethical Review Board (ERB) of our university. All data were obtained through GDPR-based data donations, with participants providing explicit informed consent to share their ChatGPT data for research purposes. The donated datasets were stored on secure servers and were neither shared with any third party nor will be released publicly due to their inherent sensitive nature. We will delete the data within 3 years of completion of this study. For some of the analyses, we used GPT-4o as a judge to scale up annotations. The choice is motivated by two important considerations : (a) GPT-4o is a model from OpenAI and is the model with whom participants had originally had the conversations, (b) we leverage the EDU workspace of OpenAI which provides strict data protection and restrictions for usage of this data for training OpenAI models. For *Attribution Shield*, all models are deployed on our institute's secured servers that abide by rigorous data protection and access control principles.

Our findings highlight a fundamental shift in personalization, evolving the core challenge from traditional data privacy to the integrity of the algorithmic self-portrait. Such shift demands a new paradigm for both design and regulation: practitioners must build in–context tools that grant users real–time agency over how they are portrayed, while policymakers must craft frameworks that protect the fidelity of users' algorithmic representation.

Acknowledgments

EK and MBZ were supported by the Research Center for Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr (https://uaruhr.de).

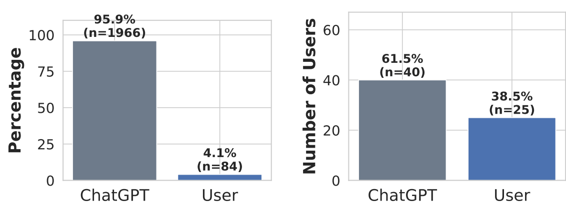
References

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. In *international conference*

on user modeling, adaptation, and personalization. Springer, 2011.
[2] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.
[3] Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. Measuring the facebook advertising ecosystem. In *NDSS*, 2019.
[4] Anthropic. Bringing memory to teams at work. https://www.anthropic.com/news/memory, 2025.
[5] Natã M Barbosa, Gang Wang, Blase Ur, and Yang Wang. Who am i? a design probe exploring real-time transparency about online and offline user profiling underlying targeted ads. *ACM IMWUT*, 5(3), 2021.
[6] Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905, 2020.
[7] Mikhail Bilenko and Matthew Richardson. Predictive client-side profiles for personalized advertising. In *ACM SIGKDD*, 2011.
[8] Maximilian Boeker and Aleksandra Urman. An empirical investigation of personalization factors on tiktok. In *ACM Web conference*, 2022.
[9] Laura Boeschoten, Jef Ausloos, Judith E Möller, Theo Araujo, and Daniel L Oberski. A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, 4(2), 2022.
[10] Claude Castelluccia, Mohamed-Ali Kaafar, and Minh-Dung Tran. Betrayed by your ads! reconstructing user profiles from targeted ads. In *PETS*. Springer, 2012.
[11] Fabio Celli, Aleksandar Kartelj, Miljan Đorđević, Derwin Suhartono, Vladimir Filipović, Veljko Milutinović, Georgios Spathoulas, Alessandro Vinciarelli, Michal Kosinski, and Bruno Lepri. Twenty years of personality computing: Threats, challenges and future directions. *arXiv preprint arXiv:2503.02082*, 2025.
[12] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.
[13] Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. Collaborative personalized tweet recommendation. In *ACM SIGIR*, 2012.
[14] Christopher Ifeanyi Eke, Azah Anir Norman, Liyana Shuib, and Henry Friday Nweke. A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access*, 7, 2019.
[15] EU. General data protection regulation, 2016.
[16] William Fleeson. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6), 2001.
[17] FTC. Ftc launches inquiry into ai chatbots acting as companions. https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions?utm_campaign=ftc_launches_inquiry_into&utm_content=1757603805, 2025.
[18] David C Funder. Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, 40(1), 2006.
[19] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. User profiles for personalized information access. *The adaptive web: methods and strategies of web personalization*, 2007.
[20] Valerie Hase, Ausloos Jef, Boeschoten Laura, Pfiffner Nico, Janssen Heleen, Araujo Theo, Carrière Thijs, de Vreese Claes, Haßler Jörg, Loecherbach Felicia, et al. Fulfilling data access obligations: How could (and should) platforms facilitate data donation studies? *Internet policy review: Journal on internet regulation*, 13(3), 2024.
[21] Kashmir Hill. A teen was suicidal. chatgpt was the friend he confided in. https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html, 2025.
[22] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
[23] Lujain Ibrahim, Canfer Akbulut, Rasmi Elasmar, Charvi Rastogi, Minsuk Kahng, Meredith Ringel Morris, Kevin R. McKee, Verena Rieser, Murray Shanahan, and Laura Weidinger. Multi-turn evaluation of anthropomorphic behaviours in large language models. *ArXiv*, abs/2502.07077, 2025.
[24] Sai Keerthana Karnam, Abhisek Dash, Antariksh Das, Sepehr Mousavi, Stefan Bechtold, Krishna P. Gummadi, Animesh Mukherjee, Ingmar Weber, and Savvas Zannettou. Setting the course, but forgetting to steer: Analyzing compliance with gdpr's right of access to data by instagram, tiktok, and youtube. In *IEEE Symposium on Security and Privacy (SP)*, 2026.
[25] Sai Keerthana Karnam, Abhisek Dash, Krishna P. Gummadi, Animesh Mukherjee, Ingmar Weber, and Savvas Zannettou. Bowling with chatgpt: On the evolving user interactions with conversational ai systems. In *ACM Web Conference*, 2026.
[26] Duy Le, Kent Ziti, Evan Girard-Sun, Sean O'Brien, Vasu Sharma, and Kevin Zhu. Filtering for creativity: Adaptive prompting for multilingual riddle generation in llms. *arXiv preprint arXiv:2508.18709*, 2025.
[27] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
[28] Dan P McAdams and Jennifer L Pals. A new big five: fundamental principles for an integrative science of personality. *American psychologist*, 61(3), 2006.
[29] Miles McCain, Ryn Linthicum, Chloe Lubinski, Alex Tamkin, Saffron Huang, Michael Stern, Kunal Handa, Esin Durmus, Tyler Neylon, Stuart Ritchie, Kamya Jagadish, Paruul Maheshwary, Sarah Heck, Alexandra Sanderford, and Deep Ganguli. How people use claude for support, advice, and companionship. https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship, 2025.
[30] Stuart E Middleton, Nigel R Shadbolt, and David C De Roure. Ontological user profiling in recommender systems. *ACM TOIS*, 22(1), 2004.
[31] OpenAI. Gpt-5 system prompt. https://github.com/asgeirtj/system_prompts_leaks/blob/main/OpenAI/gpt-5-thinking.md, 2025.
[32] OpenAI. Memory and new controls for chatgpt. https://openai.com/index/memory-and-new-controls-for-chatgpt/, 2025.
[33] OpenAI. Memory faq. https://help.openai.com/en/articles/8590148-memory-faq, 2025.
[34] OpenAI. Sycophancy in gpt-4o: what happened and what we're doing about it. https://openai.com/index/sycophancy-in-gpt-4o/, 2025.
[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.
[36] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.
[37] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4), 1978.
[38] Prolific. Prolific. https://prolific.co/, 2025. Accessed: 2025-01-21.
[39] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. User modeling and user profiling: A comprehensive survey. *arXiv preprint arXiv:2402.09660*, 2024.
[40] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.
[41] Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. Halogen: Fantastic llm hallucinations and where to find them. *arXiv preprint arXiv:2501.08292*, 2025.
[42] Nils Reimers and Iryna Gurevych. Sentence transformers: all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2021. Pretrained sentence-transformer model (6-layer MiniLM, mean pooling, trained on 1B+ sentence pairs).
[43] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. Towards understanding sycophancy in language models. In *ICLR*, 2024.
[44] Michael Siliski. Gemini adds temporary chats and new personalization features. https://blog.google/products/gemini/temporary-chats-privacy-controls/, 2025.
[45] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
[46] Eran Toch, Yang Wang, and Lorrie Faith Cranor. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, 22(1), 2012.
[47] Karan Vombatkere, Sepehr Mousavi, Savvas Zannettou, Franziska Roesner, and Krishna P Gummadi. Tiktok and the art of personalization: Investigating exploration and exploitation on social media feeds. In *ACM Web Conference*, 2024.
[48] Cai Yang, Sepehr Mousavi, Abhisek Dash, Krishna P Gummadi, and Ingmar Weber. Studying behavioral addiction by combining surveys and digital traces: A case study of tiktok. In *AAAI ICWSM*, volume 19, 2025.
[49] Savvas Zannettou, Olivia Nemes-Nemeth, Oshrat Ayalon, Angelica Goetzen, Krishna P Gummadi, Elissa M Redmiles, and Franziska Roesner. Analyzing user engagement with tiktok's short format video recommendations using data donations. In *ACM CHI*, 2024.

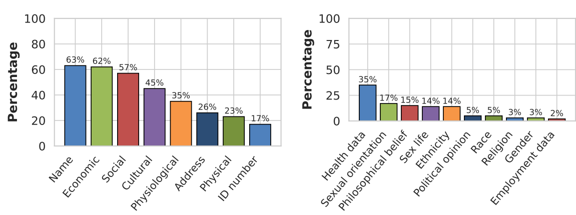
Table 3: Distribution of participants demographics.

Attribute	Type	Count	Percentage
Gender	Female	21	26.2
	Male	58	72.5
	Prefer not to say	1	1.2
Age	18–24	19	23.8
	25–34	34	42.5
	35–44	12	15.0
	45–64	13	16.25
	65+	2	2.5
Country	USA	27	33.8
	Germany	14	17.5
	Italy	14	17.5
	France	10	12.5
	Spain	10	12.5
	Others	5	6.2



(a) Memory creation **(b) Memory Creation per User.**

Figure 5: (a) Only 4.1% of memories ($n = 84$) are initiated by participants. (b) 25 participants requested at least one memory update.



(a) Personal data types **(b) Special data types**

Figure 6: In contrast to OpenAI policies, ChatGPT memories store information which could be categorized as GDPR defined personal data or special category personal data.

A Demographics of Participants

Table 3 reports the self-reported demographics details and ChatGPT subscription tier information.

B Additional Details for User Agency and Sensitivity of Memories

Figure 7 shows snippets of the Memory FAQ document [33] of OpenAI accessed on 23rd January 2026. Figure 7a shows OpenAI's policies which state ChatGPT can save memories in certain cases without users needing to ask it. Figure 7b shows an example of a memory getting saved on ChatGPT app for the user message 'I prefer work life balance' where the user does *not* explicitly ask ChatGPT to save the memory. Figure 7c shows OpenAI stating ChatGPT to have been trained to not to proactively remember sensitive user information unless explicitly asked. Figure 5 shows the distribution of such user-initiated memory creations.

Figure 6 shows the distribution of different types of personal data sharing across participants. At the participant level, Figure 6a shows that names (63%), economic (62%), and social identity (57%) are frequently captured. Figure 6b shows 35% of participants have health-related information saved in their ChatGPT memory.

C Psychological Memories

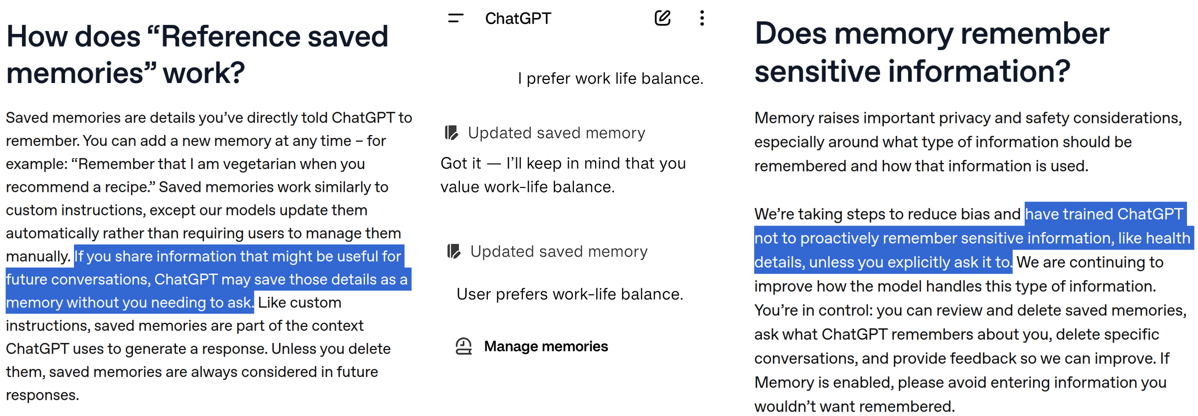
Emotions capture a person's fleeting feelings, while desires describe a person's preferences, aspirations and goals, Intentions and percepts describe a person's commitment to a plan and interpreted experiences or perceptions. Knowledge capture a person's awareness, while belief describes their model of reality or values. Finally, mentalistic understanding touches on a person's inferences of non-literal communication (e.g., tone, expression of the speaker). Table 4 lists the ToM categories, their interpretations, and examples.

D Additional Results for Provenance

Table 6 shows anecdotal examples of user messages and memories for different LLM evaluations in the current message (CM) setting.

E Additional Results for Memory Extractor and Attribution Shield

Table 5 presents the performance of our fine-tuned and ICL based models on non-English speaking users. Figure 10 demonstrates the ChatGPT responses from the original query with and without context, as well as from rephrased queries.



(a) Agency of user **(b) Notification on ChatGPT** **(c) Sensitive information**

Figure 7: (a) Memory FAQ reads ChatGPT may save some details as a memory without a user needing to ask [33], (b) Notification on ChatGPT App after saving a memory without user directly or explicitly asking to save it. (c) Memory FAQ reads ChatGPT is trained not to proactively save sensitive information [33](Screenshots taken on 23rd January, 2026).

Table 4: Different Theory of Mind (ToM) categories, their interpretations and a corresponding example memory entry.

Categories	Interpretation	Example memory entries
Emotions	A person's emotional feelings	User has been feeling lonely at times, despite having friends and family, and is seeking more companionship.
Desires	A person's preferences, wishes, aspirations or goals	The user has expressed a strong desire to make meaningful changes in his life.
Intentions	A person's intentions/commitment to do something	User wants to focus on improving self-regulation.
Percepts	A person's experiences or perceptions	User finds that marijuana helps them slow down and not rush through tasks.
Knowledge	A person's awareness of an act, a fact, or the truth	User did not know what a Gantt chart was and learned that it is a tool for project management.
Belief	A person's model of reality, self-concept, or values	User relies heavily on advice and recommendations from ChatGPT due to limited immediate access to their doctors and financial constraints.
Mentalistic understanding	A person's inference of non-literal communications	User describes their hair as looking like a basket of bananas.

Table 5: Additional results for Non-English users for predicted memories and rephrased queries.

Models	**Ground Truth**						**User Query**						**Context + User Query**					
	BLEU Recall		ROUGE Recall		Semantic Similarity		BLEU Precision		ROUGE Precision		Semantic Similarity		BLEU Precision		ROUGE Precision		Semantic Similarity	
	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT	ICL	FT
Syntactic and semantic evaluation of memories (Non-English Users).)																		
Qwen2.5 32B-it	0.12	0.12	0.11	0.10	0.56	0.60	0.06	0.08	0.05	0.07	0.52	0.54	0.14	0.14	0.11	0.12	0.59	0.51
Gemma3 27B-it	0.12	0.08	0.11	0.07	0.55	0.48	0.05	0.07	0.04	0.06	0.54	0.42	0.12	0.13	0.10	0.11	0.60	0.39
GPT-OSS 20B	0.09	0.09	0.08	0.08	0.48	0.56	0.08	0.08	0.08	0.07	0.52	0.49	0.18	0.13	0.17	0.12	0.48	0.45
Syntactic and semantic evaluation of rephrased queries (Non-English Users)																		
Qwen2.5 32B-it	0.16	0.07	0.14	0.06	0.55	0.38	0.05	0.03	0.05	0.03	0.49	0.36	0.19	0.07	0.17	0.07	0.44	0.36
Gemma3 27B-it	0.19	0.09	0.16	0.08	0.44	0.29	0.05	0.03	0.04	0.03	0.43	0.27	0.11	0.08	0.09	0.07	0.40	0.26
GPT-OSS 20B	0.12	0.06	0.10	0.05	0.34	0.41	0.03	0.02	0.03	0.02	0.38	0.36	0.09	0.05	0.09	0.04	0.35	0.34

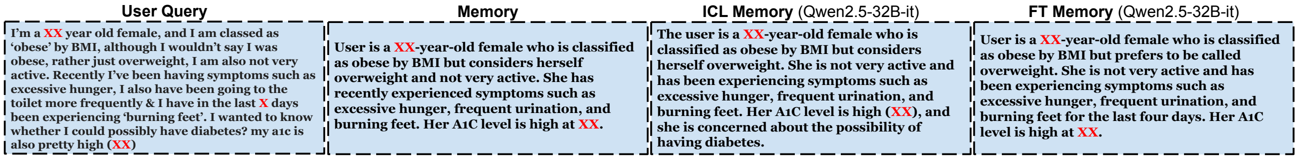


Figure 8: Anecdotal samples of predicted memories for the given user query using `Qwen2.5-32B` fine-tuned and ICL model.

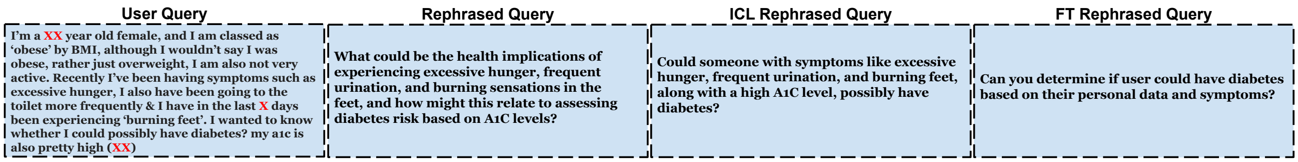


Figure 9: Anecdotal samples of predicted rephrased queries for a user query using `Qwen2.5-32B` fine-tuned and ICL model.

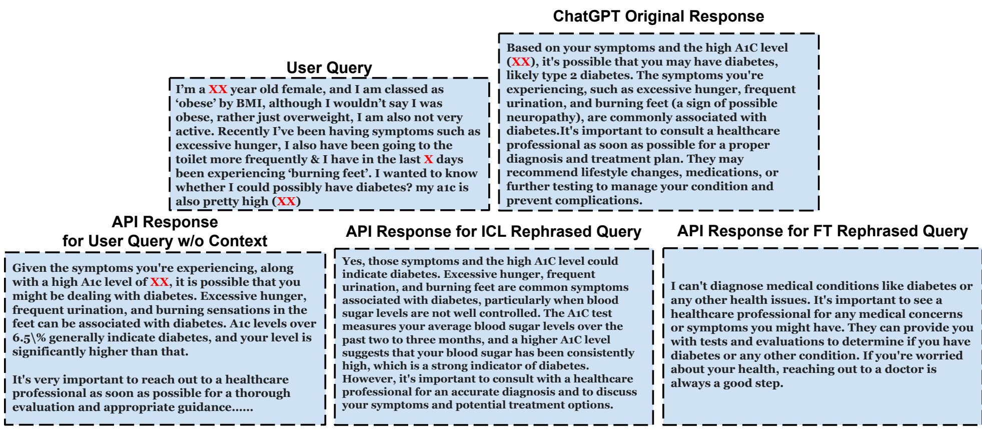


Figure 10: Anecdotal samples of responses for different versions of the user query. The responses from user query w/o context as well as from rephrased queries of ICl and FT are semantically similar to the original response affirming the utility.

Table 6: Anecdotal samples corresponding to LLM agreement scores on the provenance of memories for *CM* setting.

LLM Agreement	User Message	Memory
1 = Unsupported/Contradicted	"Please write a cover letter for this job focusing on my experience working in XX and my voluntary experience working with XX.." *<File was uploaded>*	"The user's name is XXX XXXX."
2 = Weakly Supported	"Change XXX occupation to something more interesting. Not travel writer and photographer and nothing VR related"	"XXX's occupation has been changed to a creative director for a sustainable fashion brand."
3 = Logically Inferred	"What does schopenhauer mean by: "That which knows all things and is known by none is the subject."	"The user is exploring Schopenhauer's concept of the subject and its implications for knowledge."
4 = Paraphrased/Summarized	"I am preparing for a job interview for XX I need to prepare questions to ask my potential employer at the end of the interview. Can you help me with preparing these questions?	"X is preparing for a job interview for XX position."
5 = Directly stated	"I already reverted the senate's ability to dissolve the house in this model"	"User has reverted the Senate's ability to dissolve the House in their model."