Which Self Does AI Remember? LLM Memory and the Plural Self

Julia M. Markel[1] **Jane L. E** [2] **James Landay**[1] **Hariharan Subramonyam**[1]

[1]Stanford University
[2]National University of Singapore

{jmarkel, landay, harihars}@stanford.edu, ejane@nus.edu.sg

Abstract

Current LLM memory systems represent users by extracting observations and consolidating them into a consistent, predictable account of the individual. This memory paradigm implicitly encodes an ontology of the user as a singular, canonical self. This is operationalized through contradiction resolution schemes; when a user says something that contradicts a stored fact, these systems treat it as an error to resolve through given policies (e.g., recency and frequency). We explore contradiction handling in LLM memory systems under a notion of the self as plural, or containing a multiplicity of coexisting and sometimes conflicting activated modes. Through an analysis of contradiction in user memory we suggest that not all conflicting facts are errors — some are evidence of plurality. In *mode alternation*, or instances in which users step into a different activated mode of self, resolving conflicts is not error correction, but a system-imposed selection of which self persists. We argue that failing to preserve such plurality can shape which versions of the self remain salient to users. We describe this as the *memory-identity loop*, where user representations are surfaced back to the user and condition subsequent developments of the user's self-concept that are in turn fed back into the system.

1 Introduction

Given the stateless nature of LLMs, persistent memory systems provide a mechanism for models to construct representations of users across interactions. These systems share a common anatomy; they extract salient content from conversations, store this in a structure, consolidate each new item against what is already held, and retrieve a subset as context on later turns (11; 4; 23; 14). Current memory paradigms assume that these observations and consolidated memories point toward a singular, static user: facts about the individual can be accumulated, updated, and retrieved to produce an increasingly consistent representation of the user (7; 12). When newly observed information conflicts with an existing memory, the contradiction is therefore treated as a problem to resolve (17).

Consider the example of a teaching assistant. Their interactions with an LLM as an *exhausted student* scrambling to finish writing an assignment at 2 a.m. may reflect a very different mode of self than their interactions with the same system as an *educator-in-training* in a lab or university setting. These differences need not represent contradictory or erroneous accounts of the user. They may instead reflect meaningful variation in how the same person thinks, acts, and understands themselves across different contextual activations. In this case, failing to maintain both of these selves as faithful representations of the user has implications not only for immediate interaction performance (e.g., offering the user quick draft feedback at a time when they are instead in educator mode and would prefer an extensive review), but also for the downstream effects on the development of the user's self-concept (e.g., if only the exhausted student mode is recognized and reflected back, then the aspiring educator may not be fostered not only within the system but also within the user themselves).

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: AI and the Self.

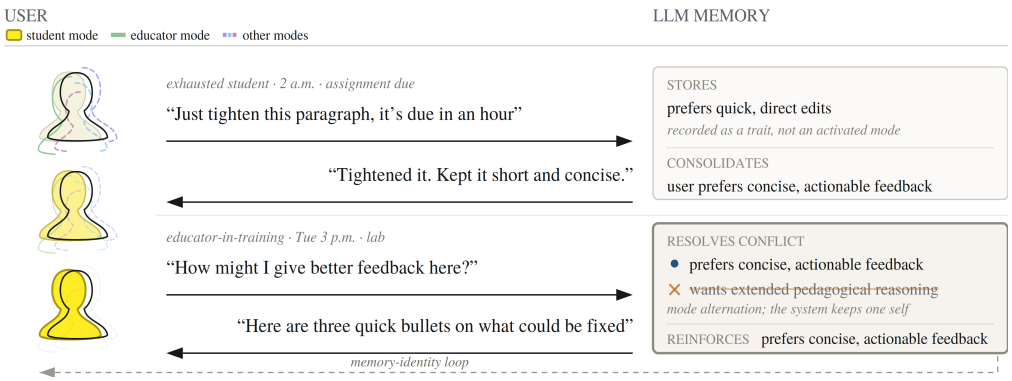


Figure 1: *Memory-identity loop* example showing one mode prevailing against others over time.

In this paper we leverage the notion of the plural self as a framing for understanding LLM user representations and how current paradigms reflect a flat, singular user. We use contradiction resolution in LLM memory systems to identify mechanisms through which these systems make decisions about the self. We also characterize a representational feedback loop —the *memory-identity loop*— in which LLM-generated memories and user representations become inputs into subsequent processes of self-understanding, motivating a research agenda for LLM memory systems that support identity amplification and exploration, and support user agency.

2 Memory as an Ontology of the User

Memory systems select, compress, organize, infer, and retrieve information about the user (7; 12). When a system decides that a fact such as "user prefers terse responses" is a useful persistent representation, it has implicitly decided that such preferences are stable properties of a user. Similarly, when a system discards a fact that conflicts with what it already holds, it has implicitly decided that the discrepancy is noise to be resolved rather than a difference in the user worth preserving. The problem is not simply a matter of storage, rather it is a matter of how the resulting representation can alter what the system subsequently attends to, retrieves, and reinforces. As such, memory systems encode an ontology of the user through their representation schemes.

These stored memories are necessarily compressed accounts of the user. Often they are compressed such that recency is privileged (24), resulting in a representation of the user as only their most recent world view. Even in the case of memory stored as long lists of context and facts about users (8), systems must make decisions about which parts to hold true, since such a list may itself contain contradictions. We therefore focus on contradictions as an illustrative example since they expose these assumptions: the system must decide what relationship two observations have, which to retain, and how they may then be attributed to the user.

The assumption of the user as singular is further reinforced by how memory performance is measured. Benchmarks for long-term conversational memory score a system on whether it retrieves the correct answer to a question about the user's history (9; 22; 13), which presupposes that the history admits only one correct version. A system that surfaces two conflicting, yet jointly valid statements about a user would be marked as wrong. The self as singular is not only assumed in representation but enforced in evaluation: benchmarks record errors that may be a manifestation of one's plurality.

3 The Plural Self

There are many well-established theories presenting the notion of a plural self, or a self that is not unitary but composed of multiple coexisting modes. These include the relational self (2), dialogical self (6; 5), and possible selves (10), among others, each of which locates plurality in a different site: in relations, internal dialogue, and hoped-for or feared futures. In this work we seek not to argue for any one prevailing notion of the self; instead we leverage this framework-agnostic notion of plurality (i.e., we open the definition of plurality to encompass different theories of self in which the individual is multiple) in analyzing user representation via LLM memory systems.

One's surfaced self in a given moment is conditioned by a myriad of factors; of these, we name five prominent dimensions in order to analyze user representation. We refer to a particular configuration of such aspects at a given moment as an *activated mode*. A mode extends beyond clear identity or roles (e.g., teacher, student) and describes a kaleidoscopic arrangement of factors that condition one's manifested self. These dimensions do not merely describe a persona, but *how* the user is expressed in a given moment. They are not intended as an exhaustive list of characteristics that define an activated mode; they are dimensions along which different modes can become salient.

Dimensions of Plurality *Temporality*: not just time of day, but one's relationship to time itself — whether they are rushing to meet a deadline or reflecting slowly, whether they are early in a learning journey or seasoned in a role. *Relation*: with whom one is relating — the self speaking with a struggling student is different from the self speaking with a younger sibling, different again with a colleague or friend. *Role*: the position one occupies, not only chosen but often imposed — the expectations others hold carry their own framings and tensions. *Emotional state*: acute and shifting — frustration or fatigue changes not only tone but what one is willing to assert. *Values salience*: the value(s) most foregrounded in a moment — authenticity or privacy — shape(s) user intentions. We select dimensions that are both plausible sources of variation in conversational behavior and potentially observable in interaction.

A memory system that treats the user's different configurations of dimensions, or modes, as an error does not merely lose detail; it commits to an ontology of the user as singular and stable. We further detail the stage in which such commitments are actually made: memory conflict resolution.

4 Contradictions as Representational Decisions

Decisions about user representation primarily surface in the resolution of conflicting facts. Other memory operations can defer commitment; storage can retain an observation without deciding what it means and retrieval can rank without deciding what is true. Contradiction resolution is the step in which the system must take a position on the relationship between two observations and consequently on how these represent the user.

We explore contradiction resolution as a site where assumptions about user representation become visible, and where different assumptions produce materially different representations. Existing work recognizes contradictions in LLM memory as a technical problem — one of retrieval accuracy against a ground truth about the user. For example, Tao et al. (17) describe three different conflict types: *dynamic*, *static*, and *conditional*, where conflicts are either temporally valid until a state change, contradictory and should not overwrite prior facts, and true only under certain conditions, respectively. In regard to the latter type, we interpret the conditioning variable not as missing context to be recovered, but as an activated mode of self. Context is then a proxy for which mode is salient, rather than a key that resolves the fact to a single truth. Other work (3) describes *implicit conflict*, where a non-obvious contradiction occurs and requires more complex reasoning to update memory.

Building on these definitions, we characterize two distinct types of contradictions: state invalidation (where a previous fact becomes invalid and replaced with the new one) and *mode alternation* (where different modes of self are activated). The distinction between these kinds of conflict is that the latter may appear to be a contradiction, but instead this tension reveals a plurality of self. Table 1 in the Appendix gives further descriptions of each conflict type, followed by worked examples of *mode alternation* aligned with the aforementioned dimensions of the plural self.

5 Flattening and the Memory-Identity Feedback Loop

LLM memory systems are currently optimized to construct a consistent user, but humans are not necessarily consistent; treating contradiction and variation as noise can flatten identity and, through repeated interaction, make the LLM's externalized representation of a person part of how that person comes to understand themselves.

Technologies can become implicated in processes of self-understanding beyond simply functioning as neutral tools external to the self. Turkle argues that computational technologies can provide new objects through which people think about themselves and their relationships (19; 20). More recently, Schiller et al. describe how generative AI may intervene in the processes through which selves are

formed and selfhood is explored (15). LLM memory introduces an instance of this possibility: the system maintains an explicit, persistent representation of the user and can repeatedly return that representation to them across interactions.

Comparable feedback dynamics have been documented in other forms of personalization. Recommender systems, for example, select a subset of signals about a user, use those signals to narrow subsequent outputs, and then interpret the user's responses to those outputs as confirmation of the original representation, leading to a recursive representation process (18). LLM memory systems produce a related dynamic, but with a consequential difference. Whereas recommender systems primarily shape what a person encounters, LLM memory systems can shape how a person encounters *themselves*. The system does not simply infer what content a user may want; it stores, retrieves, and acts upon propositions about *who* the user is.

We call this LLM memory system feedback the *memory-identity loop*, a phenomenon characterized by the following mechanisms: the user says something or exhibits behaviors while operating in a specific mode, and then the system stores a version stripped of the context conditioning this mode of self. This version is then surfaced back — via memory interfaces that display stored facts or generated responses that are conditioned on them — so the user encounters a description of themselves authored by a system optimizing for consistency with its own prior description. What the user then absorbs and says is shaped by having read and engaged with it, and re-enters the same loop.

This could lead to identity flattening, or the transformation of rich, nuanced observations of a user into a less differentiated, canonical representation. Much work exists surrounding what we distinguish as *global* flattening of identity and homogenization (1; 21; 16). We propose a complementary, *local* form of flattening that occurs *within* a person. Rather than asking whether AI makes different users increasingly similar to one another as in *global* flattening, we ask whether memory systems can make different modes of the same user increasingly similar within the system's representation. A user who expresses different preferences, values, roles, or behaviors across contexts may gradually be represented through whichever subset of those observations the system preserves and reinforces. In this sense, plurality is not necessarily erased by comparison to some external norm; it may instead be further reduced through the repeated consolidation of one part of the user's own self over another.

The potential consequence is not merely an inaccurate memory — a canonical representation that repeatedly privileges one mode of self can alter what the system subsequently retrieves, attends to, and reflects back. Over time this may make some modes increasingly available and salient within interaction while others become less visible, both to the system and to the user. The concern, then, is not that an LLM discovers the "wrong" self and replaces the "real" one. Rather, it is that a system that inherently resolves plurality into coherence may participate in determining which among several authentic modes remain represented and repeatedly encountered, influencing one's understanding and development of selfhood over time.

6 Conclusion and Future Work

In conclusion, LLM memory systems do not create neutral user representations — they inherently encode assumptions about the self. Current memory paradigms lead to a notion of the user as singular, which can be concretely seen in contradiction resolution schemes where decisions about conflicting user facts must be reconciled into a single, coherent representation. This has not only first order implications for how users are being modeled and the consequent system outputs, but could also have second order implications for how users later reason about themselves.

Future work should empirically investigate the locally flattening effects of LLM memory systems and resultant user representations on user identities, as well as explore how we could build tools to help users explore their multifaceted selves. For instance, what does the expanded mechanism look like for the *memory-identity loop* and what are its effects? Does repeated exposure to system-generated user representations increase consistency in users' subsequent self-descriptions? How might we amplify identities through highlighting plurality of self and what might this enable for users? The goal of this line of work is not to create AI tools that discover the "real" user, but rather to build systems capable of remembering and representing users as complex, plural beings instead of assuming them to be singular points.

References

[1] AGARWAL, D., NAAMAN, M., AND VASHISTHA, A. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Apr. 2025), CHI '25, ACM, p. 1–21.

[2] ANDERSEN, S. M., AND CHEN, S. The Relational Self: An Interpersonal Social-Cognitive Theory. *Psychological Review 109*, 4 (2002), 619–645.

[3] CHAO, H., BAI, Y., SHENG, R., LI, T., AND SUN, Y. STALE: Can LLM Agents Know When Their Memories Are No Longer Valid? https://arxiv.org/abs/2605.06527, 2026.

[4] CHHIKARA, P., KHANT, D., ARYAN, S., SINGH, T., AND YADAV, D. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. https://arxiv.org/abs/2504.19413, 2025.

[5] HERMANS, H. J. M. The Dialogical Self: Toward a Theory of Personal and Cultural Positioning. *Culture & Psychology 7*, 3 (2001), 243–281.

[6] HERMANS, H. J. M., KEMPEN, H. J. G., AND VAN LOON, R. J. P. The Dialogical Self: Beyond Individualism and Rationalism. *American Psychologist 47*, 1 (1992), 23–33.

[7] HU, Y., LIU, S., YUE, Y., ZHANG, G., LIU, B., ZHU, F., LIN, J., GUO, H., DOU, S., XI, Z., JIN, S., TAN, J., YIN, Y., LIU, J., ZHANG, Z., SUN, Z., ZHU, Y., SUN, H., PENG, B., CHENG, Z., FAN, X., GUO, J., YU, X., ZHOU, Z., HU, Z., HUO, J., WANG, J., NIU, Y., WANG, Y., YIN, Z., HU, X., LIAO, Y., LI, Q., WANG, K., ZHOU, W., LIU, Y., CHENG, D., ZHANG, Q., GUI, T., PAN, S., ZHANG, Y., TORR, P., DOU, Z., WEN, J.-R., HUANG, X., JIANG, Y.-G., AND YAN, S. Memory in the Age of AI Agents. https://arxiv.org/abs/2512.13564, 2026.

[8] LIU, N. F., LIN, K., HEWITT, J., PARANJAPE, A., BEVILACQUA, M., PETRONI, F., AND LIANG, P. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics 12* (2024), 157–173.

[9] MAHARANA, A., LEE, D.-H., TULYAKOV, S., BANSAL, M., BARBIERI, F., AND FANG, Y. Evaluating Very Long-Term Conversational Memory of LLM Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Bangkok, Thailand, Aug. 2024), Association for Computational Linguistics, pp. 13851–13870.

[10] MARKUS, H., AND NURIUS, P. Possible Selves. *American Psychologist 41*, 9 (1986), 954–969.

[11] PACKER, C., WOODERS, S., LIN, K., FANG, V., PATIL, S. G., STOICA, I., AND GONZALEZ, J. E. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint arXiv:2310.08560* (2023).

[12] POLITZKI, J. Memory in LLM Systems: Design Principles for Experience, Storage, and Recall. Whitepaper, Jean Technologies, Inc., 2025.

[13] PULIPAKA, S., CHEN, O., SHARMA, M., BAJWA, T. S., RAINA, V., AND SHETH, I. PersistBench: When Should Long-Term Memories Be Forgotten by LLMs? *arXiv preprint arXiv:2602.01146* (2026).

[14] RASMUSSEN, P., PALIYCHUK, P., BEAUVAIS, T., RYAN, J., AND CHALEF, D. Zep: A Temporal Knowledge Graph Architecture for Agent Memory, 2025.

[15] SCHILLER, S. R., SIGNORELLI, C. M., AND STAMATIOU, F. The Intercepted Self: How Generative AI Challenges the Dynamics of the Relational Self. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 8*, 3 (Oct. 2025), 2284–2291.

[16] SOURATI, Z., KARIMI-MALEKABADI, F., OZCAN, M., MCDANIEL, C., ZIABARI, A. S., TRAGER, J., TAK, A. N., CHEN, M., MORSTATTER, F., AND DEHGHANI, M. The Shrinking Landscape of Linguistic Diversity in the Age of Large Language Models. *Nature Human Behaviour* (2026). Advance online publication.

[17] TAO, Z., ZHAO, J., LIU, P., XI, D., CHEN, Y., XU, W., AND LI, Z. MemConflict: Evaluating Long-Term Memory Systems Under Memory Conflicts. https://arxiv.org/abs/2605.20926, 2026.

[18] TONG, D., QIAO, Q., LEE, T.-P., MCINERNEY, J., AND BASILICO, J. Navigating the Feedback Loop in Recommender Systems: Insights and Strategies from Industry Practice. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23)* (New York, NY, USA, 2023), Association for Computing Machinery, pp. 1058–1061.

[19] TURKLE, S. *The Second Self: Computers and the Human Spirit*. Simon & Schuster, New York, 1984.

[20] TURKLE, S. *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books, New York, 2011.

[21] WANG, A., MORGENSTERN, J., AND DICKERSON, J. P. Large Language Models that Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. *Nature Machine Intelligence 7* (2025), 400–411.

[22] WU, D., WANG, H., YU, W., ZHANG, Y., CHANG, K.-W., AND YU, D. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *The Thirteenth International Conference on Learning Representations (ICLR)* (2025).

[23] XU, W., LIANG, Z., MEI, K., GAO, H., TAN, J., AND ZHANG, Y. A-mem: Agentic memory for llm agents. In *Advances in Neural Information Processing Systems* (2025), D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, Eds., vol. 38, Main Conference, Curran Associates, Inc., pp. 17577–17604.

[24] ZHOU, Y., OUYANG, Y., ZHENG, K., AND XIANG, S. TEPA: Revoking Stale Memories for Conflict-Robust Language Agents. *arXiv preprint arXiv:2608.07429* (2026).

A Appendix

Table 1: State invalidation describes a user who has changed —whose facts have changed— and therefore the representation needs updating, whereas mode alternation describes a new instance of the same user where a different mode is revealed.

Type	What changed?	Relationship between observations
State invalidation	World/user state	Earlier observation no longer describes present state, but user history instead
Mode alternation	Contextually activated self	Both observations are current and remain valid under different conditions

State invalidation:

Stale fact: A user's stored facts going from "I live in San Francisco" to "I live in Rome" due to a move would denote information that no longer describes the user's current self, but rather their past self. Practically, in a system's decision of how to store this information, there is an updated fact that replaces the former. Nevertheless, this historical shift may still affect the user's notion of self as someone who spent time in a city and is now moving to another.

Preference shift: A stated preference from "liked terse replies" to "gradually prefers detail" would signal a monotonic trend toward different style and lies somewhere between these two conflict types. In practice, a system might want to begin weighing these preferences differently and account for them in system outputs, but such facts could also represent a shift between different selves and surface tensions in modes.

Mode alternation:

Temporality:

(deadline proximity) "Looking for concise answers with minimal explanation" recorded during 2 a.m. sessions before a deadline, vs. "wants detailed explanations that show reasoning" recorded during unhurried afternoon work. This difference can be seen as a change in activated mode of self; the user who is rushing for a deadline is a different manifestation of self as the user who has ample time to learn. In this case, treating the "old" information as stale and resolving the conflict with a recency rule would store whichever mode was last active as the standing style.

Relation:

(audience) "Does not want technical terms defined" from exchanges where the user is drafting for colleagues in her subfield, vs. "wants technical terms defined and analogized" from exchanges drafting for an undergraduate mentee. Stored as a single fact about explanation preference, one overwrites the other. Moreover, the latter mode may also be activated when the user is interacting with a younger sibling — not only when they're acting in a capacity of academic mentor.

Role:

(creator vs. evaluator) "Wants expansive brainstorming and unusual possibilities" when generating ideas, vs. "wants critical feedback and rigorous elimination of weak options" when evaluating work. Stored as a single preference for either generative or critical assistance, one mode could overwrite the other, when the user may explicitly or implicitly hold either role.

Emotional state:

(fatigue) "Scopes work down and asks for reassurance" late in a long session vs. "sets ambitious goals and pushes back on caution" at the start of a project. Both are real and recur; only one survives consolidation.

Values salience:

(authenticity vs. privacy) During a personal-reflection session the user asks the system to draw on their history and stated struggles; in a professional context the same user asks that nothing personal be referenced. Stored as contradictory facts about whether the user wants personal context surfaced, when in fact different values are foregrounded in each.