What Yang keeps

Memory in conversational agents: state of the art, a reading of *After Yang* and a tagging proposal with Mem0

Cris Olmedo and Kurt Sommer
Divina Machina Research Group
Working paper · State of the art
DM-2026-10, version 1.0 · 1 November 2026

Abstract. Conversational agents are beginning to remember. In two years, memory has gone from a long-context trick to a product layer with its own architecture, benchmarks and privacy problems. This paper reviews that state of the art: the reference systems (Generative Agents, MemGPT, Mem0, Zep, A-MEM), the benchmarks used to measure them (LoCoMo, LongMemEval) and what those figures do and do not allow us to conclude.

We read it alongside *After Yang* (Kogonada, 2021), a film about the memory of a machine that nobody knew existed, discovered once the machine has stopped working. The film raises the questions benchmarks do not measure: what is kept, who appears in what is kept without having agreed to it, whose memory it is when the agent is gone and who decides to share it.

We propose a tagging scheme for Mem0 designed for a public character such as GAIA 2147, which does not remember people, and for a personal companion, which could do so with permission. We close with a foresight exercise on how we believe agent memory will evolve. The work is a targeted review without own data; the scheme and the foresight are hypotheses. The companion paper DM-2026-11 deals with the custody and encryption of that memory.

Keywords: agent memory; Mem0; LoCoMo; LongMemEval; After Yang; tagging; privacy; foresight.

1. Introduction

An agent without memory starts every conversation from scratch. An agent with memory can pick up a thread, remember a preference or mistake one person for another. Between those extremes lie design decisions that until recently were made by infrastructure, by accident: what is extracted from a conversation, how it is summarised, when it is retrieved, when it is forgotten and who can see it.

Our interest is practical. GAIA 2147, the character who answers in a telephone booth in public spaces, remembers nobody between calls: that was a decision, because a public booth receives children and strangers who did not agree to be remembered [1]. At the same time, a memory of the place or of the character itself could make it more continuous. We want to use Mem0, a memory layer for agents, and we need to know what the state of the art offers, what its figures really measure and how to tag what we keep.

1.1. Contributions

1. A review of the state of the art in agent memory, with a taxonomy and a comparative table of systems.
2. A critical reading of the results published on LoCoMo and LongMemEval, separating what they measure from what is usually concluded.
3. A reading of *After Yang* as a design case: archive, consent, ownership and legacy of a machine's memory.
4. A tagging scheme for Mem0 in two scenarios, public character and personal companion, marked as a hypothesis.
5. A foresight exercise with verifiable signals and horizons declared as conjecture.

1.2. Scope

The review is targeted, not systematic. Peer-reviewed articles, preprints from the groups that designed each system and official documentation are prioritised. Several results come from the developers themselves and are marked as such [FAB]. The cut-off date is 1 November 2026; the field moves fast and some figures may already be outdated.

2. What remembering is

Human memory is not an archive. In 1885 Ebbinghaus described how what is learned is lost quickly at first and slowly afterwards [2]; a common way of stylising that curve is

$$R(t) = e^{-t/S}, \tag{1}$$

where R is retention, t the elapsed time and S the stability of the memory [DEF]. It is an idealisation, not a law with universal parameters. Bartlett showed that remembering is reconstructing: people fill in, simplify and adapt what they remember to their schemas [3]. Tulving separated episodic memory, of events situated in time, from semantic memory, of facts and meanings [4]. And research on reconsolidation found, in rats and with conditioned fear memories, that retrieving a memory makes it unstable and requires it to be consolidated again [5] [EMP]. Remembering changes what is remembered.

An agent's memory does almost the opposite: it stores a version, retrieves it intact and, unless something else is designed, never forgets it. That difference is the core of this work [INT]. An agent that remembers like an archive can be more accurate than a person

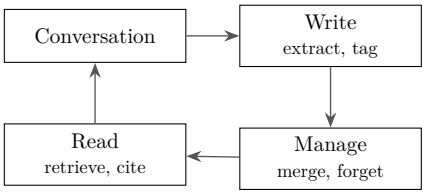


Figure 1. Memory loop of an agent, adapted from Du [7]. Each step is a design decision.

and, at the same time, less able to let someone change. Paper DM-2026-01 frames it as a revisable memory, presented with provenance and open to correction [6].

3. State of the art in agent memory

3.1. Taxonomy

Du organises the field as a write–manage–read loop, coupled to perception and action, and proposes three axes: temporal scope (how long a memory lasts), substrate (the form in which it is stored) and control policy (who decides the operations) [7]. Figure 1 adapts that loop.

3.2. Reference systems

Generative Agents. Park and colleagues gave simulated agents a natural-language memory stream from which memories were retrieved with a score that adds recency, importance and relevance [8]:

$$s(m) = \alpha_r \,\hat{r}(m) + \alpha_i \,\hat{\imath}(m) + \alpha_v \,\hat{v}(m), \qquad r(m) = 0.995^{\,\Delta h}, \tag{2}$$

where recency decays per simulated hour since last access, importance is a 1–10 rating assigned by the model when the memory is created, relevance is the cosine similarity with the query, the three terms are normalised to $[0, 1]$ and the weights equal 1. When the accumulated importance of recent events exceeds 150, the agent reflects and writes higher-level memories [EMP].

MemGPT. Packer and colleagues treat the model's context as the main memory of an operating system and move information between fast and slow tiers to simulate a larger context [9].

Mem0. It extracts the relevant facts from each conversation, consolidates them with existing ones and retrieves them when needed; a variant adds a graph of relations [10]. Each memory is associated with a user and, on the platform, receives categories assigned by a classifier [11].

Zep. It builds a temporally aware knowledge graph (Graphiti) that combines conversation and structured data and preserves the history of relationships [12].

A-MEM. Inspired by the Zettelkasten method, each memory is a note with a description, keywords and tags that is linked to related notes; adding a note can update earlier ones, which the authors call memory evolution [13].

Products. In 2025 ChatGPT extended its memory to take past conversations into account, with controls to turn it off [14]; Anthropic extended Claude's memory in 2026 [15].

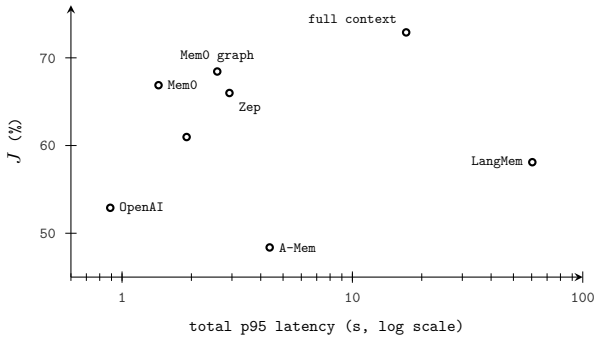


Figure 2. Quality versus latency on LoCoMo, using the data in Table 2. Full context wins on quality and loses on time and cost.

4. What the benchmarks measure

LoCoMo gathers very long dialogues, about 300 turns and 9000 tokens on average, over up to 35 sessions, generated by agents with personas and event graphs and reviewed by annotators; it evaluates question answering, event summarisation and multimodal dialogue, and finds that models fall well short of humans [16] [EMP]. *LongMemEval* measures five abilities (information extraction, multi-session reasoning, temporal reasoning, knowledge updates and abstention) with 500 questions, and observes an accuracy drop of around 30 % in commercial assistants and long-context models over sustained interactions [17] [EMP].

Table 2 reproduces the LoCoMo results published by the Mem0 authors, and Figure 2 places them on the quality–latency plane.

Four caveats before using these figures [INT]:

1. *Who measures.* The authors evaluate their own system with their own configuration [FAB].
2. *How it is judged.* The judge is another model, instructed to be generous: an answer counts as correct if it touches the same topic or period [10].
3. *Uneven comparisons.* OpenAI's memory was tested by pasting conversations into ChatGPT and was given all generated memories; many lacked dates, which hurt temporal questions. Zep improved if one waited hours after ingestion [10].
4. *Synthetic data.* LoCoMo consists of generated and edited conversations, not real life; the adversarial category was excluded for lack of reference answers [10, 16].

What does hold: external memory greatly reduces cost and latency compared with passing the whole history, at the price of a few quality points; and no system matches humans on LoCoMo [EMP]. What does not hold: a fixed ranking among commercial systems. Zep, for example, publishes its own favourable results on other benchmarks [12].

None of these benchmarks measures what matters in a public booth: whether remembering was allowed, whether the person knew they were being remembered and whether they can ask to be forgotten.

Table 1. Reference memory systems, organised by Du's three axes [7].

System	Substrate	Control	Temporal scope	Evidence
Generative Agents	Text memory stream with scoring	Fixed (formula) and threshold-based reflections	Whole simulation	[EMP]
MemGPT	Context tiers and external store	The model itself, via function calls	Across sessions	[EMP]
Mem0	Extracted facts; graph variant	LLM extraction and consolidation	Per user, session and agent	[FAB]
Zep	Temporal knowledge graph	Service with background processing	Relationship history	[FAB]
A-MEM	Linked, tagged notes	The agent creates, links and rewrites	Evolving	[EMP]

Table 2. LoCoMo results published by the Mem0 authors [10]. *J*: percentage of answers that a judge model (gpt-4o-mini) deems correct; mean of 10 runs.

Method	*J* (%)	p95 (s)	Tokens
Full context	72.9	17.1	26 031
Mem0 with graph	68.4	2.6	3616
Mem0	66.9	1.4	1764
Zep	66.0	2.9	3911
Best RAG	61.0	1.9	n/a
LangMem	58.1	60.4	127
OpenAI memory	52.9	0.9	4437
A-Mem	48.4	4.4	2520

5. What Yang keeps

After Yang is an American film written, directed and edited by Kogonada, based on Alexander Weinstein's short story "Saying Goodbye to Yang", premiered at Cannes in 2021 [18, 19]. A family adopts Mika and buys Yang, an android who keeps her company and connects her with her Chinese heritage. When Yang stops working, a technician finds a hidden device and a museum specialist identifies it as a memory bank that records key experiences of the unit; the manufacturer had denied its existence, according to the film to avoid a privacy scandal [18]. The father goes through those fragments and discovers Ada, a young woman who appears again and again, and earlier lives of Yang with other families. In the end they decide not to display the body at the museum but to preserve and share his memories, because his existence mattered to many people [18].

Kogonada sought a memory interface that would not look like a desktop with folders and formally separated Yang's recorded memories from human memories, which are reinterpreted [20]. It is a filmmaking decision, not a thesis about the brain, but it matches the distinction in Section 2 [INT].

We read five design questions in the film [INT]:

1. *Hidden memory.* Nobody knew Yang remembered. An agent memory the user is unaware of is not a feature: it is surveillance.
2. *Selection.* The bank keeps seconds of experiences, not everything. What is kept defines the character as much as what it says; the tag comes before the data.
3. *Third parties.* Ada appears in Yang's memories without having agreed to it. Every personal memory contains other people.
4. *Ownership after loss.* When Yang is gone, the family, the museum and the technician contend over what to do with what was kept. Memory outlives the agent.
5. *Legacy.* The final decision is neither to erase nor to exhibit, but to preserve and share with those to whom Yang mattered. It is an act of custody, not of storage.

The last four are the subject of the companion paper DM-2026-11 [21].

6. How we want to tag with Mem0

This section is a design proposal, not an implementation: GAIA does not yet use Mem0 and the scheme has not been tested [HIP].

6.1. What Mem0 offers

According to its documentation, each project has 15 default categories (for example, personal details, family, travel or food) that a classifier assigns when a memory is stored. They can be replaced with custom categories, each with a name and a description that helps the classifier. A list passed in a call completely replaces the project's list for that call, and lists are not merged. Categories are assigned at ingestion: changing the list does not retag what is already stored. For fixed values that are already known, using metadata rather than categories is recommended [11] [FAB].

Two design rules follow [INT]: what the model must infer goes into categories with precise descriptions; what we know for certain (scope, consent, provenance, temperature, expiry) goes into metadata, because it must not depend on a classifier.

6.2. Proposed scheme

Table 3 summarises the scheme. In the public-character scenario, the `persona` scope does not exist: GAIA only writes `hilo` with call scope, which expires when the call ends; `tema_sala`, aggregated and anonymous; and `canon`, written by the team. It is the layered memory we proposed in DM-2026-05 [1], expressed in Mem0's vocabulary. In the personal-companion scenario, `preferencia`

Table 3. Proposed tagging scheme for Mem0. Categories are inferred by the classifier; metadata are set by the system. Tag names are kept in Spanish, as they would appear in our system.

Type	Name	Description for the classifier, or value	Scenario
Category	`hilo`	Something the person has said in this conversation that is worth picking up again (thread)	Both
Category	`tema_sala`	What people are talking about today, without data on who said it (room topic)	Public character
Category	`canon`	A fixed memory of the character itself that must not be contradicted	Public character
Category	`preferencia`	A taste or way of being addressed that the person has asked for (preference)	Personal companion
Category	`relato`	An event the person has told about their life (story)	Personal companion
Category	`tercero`	Mention of another person who is not part of the conversation (third party)	Both (handled with more caution)
Metadata	`alcance`	Scope: `llamada` (call), `sala` (room), `personaje` (character) or `persona` (person)	Both
Metadata	`consentimiento`	Reference to the authorisation that permits storing; empty if none	Both
Metadata	`procedencia`	Provenance: said by the person, inferred by the system or written by the team	Both
Metadata	`temperatura`	Temperature: `caliente` (hot), `templada` (warm) or `fría` (cold) (see DM-2026-11)	Both
Metadata	`caduca`	Date after which it is deleted or cooled (expiry)	Both

and `relato` require consent and provenance, and anything tagged `tercero` is stored in summary form or not at all.

Provenance answers Bartlett and Kogonada: distinguishing what the person said from what the system inferred is what allows a memory to be presented as revisable rather than as identity [HIP].

6.3. How we would evaluate it

Three measures before any deployment [HIP]: agreement between the classifier's categories and those of two people on a sample of conversations; the proportion of memories with incompatible scope or consent (for example, `persona` without consent), which must be zero; and, for the personal companion, whether the person recognises as their own the memories shown to them and how many they correct.

7. How we believe it will evolve

Table 4 separates the signals, which are cited, from our conjectures, which are not [HIP].

Our bet is that the advantage will not lie in remembering more, but in remembering with permission, with provenance and with an expiry date [HIP]. An agent that forgets well can be more trustworthy than one that remembers everything.

8. Threats to validity

- Targeted review: complete coverage is not claimed.
- Several figures come from the developers of each system and have not been replicated here.
- The field changes within months; the cut-off date matters. Mem0's own documentation announces algorithms later than the article, with results from its managed platform that are not comparable with those in the article.
- The reading of the film is interpretive and relies on a reference entry and an interview.
- The tagging scheme has not been implemented or measured.

9. Recommendations for the community

1. When citing a memory result, state who measured it, with which judge and against which baseline.
2. Separate in storage what is inferred by a model (categories) from what is known for certain (metadata: scope, consent, provenance, expiry).
3. Also measure abstention, updating and corrections by the person, not just accuracy.
4. Do not store memories of people in public spaces; use call scope, aggregated memory of the place and the character's canon.
5. Treat mentioned third parties as sensitive data by default.

10. Conclusions

The state of the art already knows how to extract, consolidate and retrieve memories at a reasonable cost; it does not yet know, nor measure, whether it should have remembered them. *After Yang* reminds us that a machine's memory contains other people and outlives the machine. Tagging well is the first form of care: saying what each memory is, where it comes from, who authorised it and when it should be cooled or forgotten.

Table 4. Foresight on agent memory. The signal is published; the horizon and the implication are our conjectures.

Trend	Signal	Horizon	Design implication
Memory becomes a product layer	ChatGPT and Claude memory; layers such as Mem0 and Zep [10, 12, 14, 15]	Now	Choosing what not to remember will be a brand decision
From loose facts to time-aware graphs	Mem0 graph variant; Graphiti [10, 12]	1–2 years	Memories will have dates and validity, and may become obsolete
The agent manages its own memory	A-MEM; learned control [7, 13]	1–3 years	It will be necessary to audit what the agent rewrites about a person
Multi-session evaluation	LoCoMo, LongMemEval [16, 17]	Now	Measure abstention and updating, not just accuracy
Forgetting as a feature	Recency and decay [8]; reconsolidation [5]	2–3 years	Expiry and cooling by design
Memory held in the person's custody	Encryption and social recovery in wallets [22]	3–5 years	Portable, encrypted memory with heirs (DM-2026-11)

Signed. Cris Olmedo and Kurt Sommer, Divina Machina Research Group. *How to cite*: Olmedo, C. and Sommer, K. (2026). What Yang keeps. Working paper DM-2026-10, v 1.0. Divina Machina Research Group. Divina Machina by Qs Ventures.

References

Sources consulted up to 1 November 2026. Developer results are not Divina Machina measurements.

[1] C. Olmedo and K. Sommer. Personality as a vector of habits. drift of tone rules and within-call memory in a voice agent. Working paper DM-2026-05. Divina Machina Research Group, 2026.

[2] H. Ebbinghaus. *Über das Gedächtnis*. Duncker & Humblot, Leipzig, 1885.

[3] F. C. Bartlett. *Remembering. A Study in Experimental and Social Psychology*. Cambridge University Press, 1932.

[4] E. Tulving. Episodic and semantic memory. In E. Tulving and W. Donaldson, editors, *Organization of Memory*, pages 381–403. Academic Press, 1972.

[5] K. Nader, G. E. Schafe, and J. E. LeDoux. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406:722–726, 2000. doi: 10.1038/35021052.

[6] C. Olmedo and K. Sommer. The art of conversation. from the Before trilogy to the design of conversational agents. Working paper DM-2026-01, v0.2. Divina Machina Research Group, 2026.

[7] P. Du. Memory for autonomous LLM agents. mechanisms, evaluation, and emerging frontiers. arXiv:2603.07670, 2026. URL https://arxiv.org/abs/2603.07670.

[8] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *UIST '23. 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. URL https://arxiv.org/abs/2304.03442.

[9] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez. MemGPT: Towards LLMs as operating systems. arXiv:2310.08560, 2023. URL https://arxiv.org/abs/2310.08560.

[10] P. Chhikara, D. Khant, S. Aryan, T. Singh, and D. Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. arXiv:2504.19413, 2025. URL https://arxiv.org/abs/2504.19413.

[11] Mem0. Custom categories. Mem0 platform documentation, n.d. URL https://docs.mem0.ai/platform/features/custom-categories.

[12] P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, and D. Chalef. Zep: A temporal knowledge graph architecture for agent memory. arXiv:2501.13956, 2025. URL https://arxiv.org/abs/2501.13956.

[13] W. Xu, Z. Liang, K. Mei, H. Gao, J. Tan, and Y. Zhang. A-MEM: Agentic memory for LLM agents. In *NeurIPS 2025*, 2025. URL https://arxiv.org/abs/2502.12110.

[14] OpenAI. Memory and new controls for ChatGPT, 2025. URL https://openai.com/index/memory-and-new-controls-for-chatgpt/. April 2025 update.

[15] Bloomberg. Anthropic tries to win users from ChatGPT with memory feature, March 2026. URL https://www.bloomberg.com/news/articles/2026-03-03/anthropic-tries-to-win-users-from-chatgpt-with-memory-feature.

[16] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang. Evaluating very long-term conversational memory of LLM agents. arXiv:2402.17753, 2024. URL https://arxiv.org/abs/2402.17753.

[17] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *ICLR 2025*, 2025. URL https://arxiv.org/abs/2410.10813.

[18] Wikipedia. After Yang. English-language article. URL https://en.wikipedia.org/wiki/After_Yang. Credits, cast, source material and plot.

[19] A. Weinstein. *Children of the New World*. Picador, 2016. Includes the story "Saying Goodbye to Yang".

[20] C. Aguilar. Embodiments of time: Kogonada on After Yang. RogerEbert.com, 2022. URL https://www.rogerebert.com/interviews/embodiments-of-time-kogonada-on-after-yang. Primary interview, 3 March 2022.

[21] C. Olmedo and K. Sommer. Cold memory and hot memory. custody, encryption and transcendence of personal memories. Working paper DM-2026-11. Divina Machina Research Group, 2026.

[22] V. Buterin. Why we need wide adoption of social recovery wallets. Personal blog, January 2021. URL https://vitalik.eth.limo/general/2021/01/11/recovery.html.