DEMOTION BEATS DELETION: GRADED FORGETTING FOR CAPACITY-BOUNDED AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent LLM agents accumulate memory faster than they can afford to read it, and every deployed system resolves this by deleting. Anthropic's Memory Stores cap a store at 10,000 entries and instruct operators to remove stale ones; Mem0 prunes through LLM-issued `DELETE`; ChatGPT's background consolidation rewrites in place. Deletion is an irreversible decision taken without knowing which memory a future query will need. We ask whether it is necessary. We present **graded forgetting**: traces are demoted through discrete retention stages and compressed to progressively lossier fidelity tiers, but never removed, so a capacity budget is met by shedding *detail* rather than *items*. We introduce **recall-after-eviction**, a protocol that isolates the capability at stake by conditioning on questions whose supporting evidence a capacity bound has destroyed. On an 833-day, 1,115-memory timeline held to 35% of its full size, deletion answers **2.5%** of questions about the evicted past while graded forgetting answers **60.0%** - 80% of the uncapped ceiling, from the same byte budget. Against Mem0, the strongest published baseline and one with no capacity policy of its own, the ordering inverts precisely when the budget binds: Mem0 leads without a bound and falls to **26.7%** under one, against **66.7%** for graded forgetting. The contribution is a capability under capacity pressure rather than an accuracy gain, and we argue this is the axis on which agent memory should be evaluated.

1 INTRODUCTION

Managed agent memory is deployed infrastructure. Anthropic's Memory Stores (Anthropic, 2026a) mount a persistent, versioned document collection into an agent's sandbox. Mem0 (Chhikara et al., 2025) maintains a compact fact store through LLM-driven add/update/delete. OpenAI's ChatGPT synthesises long-horizon user memory in a background pass (OpenAI, 2026b). The question is no longer whether an agent can remember, but what it should stop remembering.

Every one of these answers by deletion, and their own documentation is explicit about the consequence:

> "When a store reaches its **10,000-memory limit**, writes to new memories fail… **Delete stale or redundant memories**… Switch your sessions over to that output store, then archive or delete the original." (Anthropic, 2026a)

Deletion is irreversible and is decided without knowledge of future queries. A memory judged stale today is unavailable tomorrow at any cue strength. This is the failure mode we target.

Graded forgetting. Human memory does not delete; it fades, and faded traces return under sufficient contextual support (Proust, 1913; Aggleton & Brown, 1999; Tulving, 1983). We operationalise this as a lifecycle over an ordinary memory store: each trace carries a retention stage k that advances with disuse and a fidelity tier ℓ that rises with k, discarding episodic detail while retaining a routing scaffold. A capacity budget is met by *compressing* rather than *removing*, so every trace remains addressable, in degraded form, indefinitely.

What we measure, and what we found. Our headline result is narrow and, we believe, the right shape for this problem. Under a hard capacity bound, questions about the evicted past are answerable

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

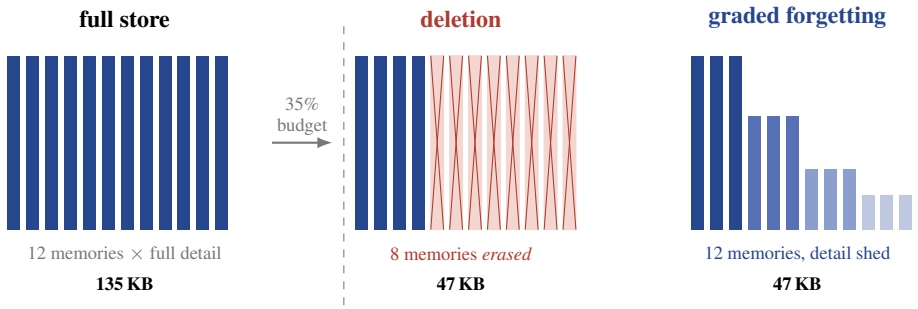


Figure 1: The two ways to meet a capacity bound. Both shed 65% of the store. Only one of them can still be asked about the material it gave up.

at 60.0% under graded forgetting and 2.5% under deletion, recovering 80% of the uncapped ceiling at identical storage cost.

We then repeat the comparison against Mem0 (Chhikara et al., 2025), which is the strongest baseline available and which, we show, has no capacity policy at all: across 381 ingestions its reconciler issued 381 `ADD` operations and zero `UPDATE` or `DELETE`, so the store grows linearly and overflows any budget. Held to the same byte cap, Mem0 must shed 73% of its memories by deletion and falls to 26.7%, while graded forgetting deletes nothing and answers 66.7%.

Contributions.

- **Recall-after-eviction** (Sec. 4), a protocol that conditions on questions whose evidence a capacity bound destroyed. We show that the obvious alternative - stratifying random questions by age - is uninformative, because evidence is duplicated across a median of eight memories and uniform eviction therefore destroys nothing.
- **A lifecycle mapped onto deployed primitives** (Sec. 3) with no new API surface: stage becomes a path prefix, demotion an ordinary update, and the platform's own version history a free recovery ladder.
- **The capability result** (Sec. 5): 60.0% vs 2.5% under a 35% budget.
- **A comparison against Mem0 under capacity pressure** (Sec. 5), showing that the strongest published baseline has no capacity policy of its own and loses two thirds of its accuracy when forced to meet a budget by deletion.

2 Background: What Deployed Memory Does

Everything is tokens. We inspected both major vendor SDKs. Neither the Anthropic Messages API nor the OpenAI Responses API has a memory role, field, or content type; retrieved memory reaches the model as ordinary tokens in the context window. Anthropic mounts a filesystem the agent reads with file tools, yielding tool-result blocks; OpenAI concatenates into the prompt. Any "memory schema" is a rendering convention honoured because models are trained on such shapes. This bounds what any memory architecture can do: the only lever is *which text enters the window*, which is precisely what a retention policy decides.

Deletion is the universal policy. Anthropic caps stores and recommends manual removal; version history exists but is reachable only through a developer-facing management API and is retained 30 days. Mem0's reconciler issues `DELETE` with no recovery path. We note an empirical observation about Mem0 in Sec. 5: on our corpus its reconciler fired *zero* times in 381 ingestions, because the conversation contains almost no contradictions - the consolidation mechanism is real but workload-dependent.

Table 1: Retention mechanisms in deployed systems. The lower block is unaddressed.

Mechanism	RAG	Mem0	Anthropic	Ours
Persistent store	-	✓	✓	✓
Consolidation	-	✓	✓	✓
Point-in-time recovery	-	-	✓ (30 d, out-of-band)	✓
Automatic ageing	-	-	-	✓
Graded accessibility	-	-	-	✓
Capacity met without deletion	-	-	-	✓
In-band recovery of evicted material	-	-	-	✓

3 METHOD

3.1 STAGE, TIER, AND THE CAPACITY POLICY

Each memory carries a retention stage advancing with time since last access,

$$k \;=\; \min\big(K,\; \lfloor \lambda \ln(1+\Delta t) \rfloor\big), \qquad p_{\text{base}}(k) = \gamma^k, \qquad \tau = \min\{k : \gamma^k < \theta\}, \tag{1}$$

and a fidelity tier $\ell(k)$ that rises with it. Tier 0 holds the verbatim episode; successive tiers discard episodic detail and retain the routing scaffold - tag, gist, topic - longest. Given a byte budget B, the policy raises tiers on the oldest material until the store fits. **No memory is removed.**

Mapped onto Anthropic Memory Stores this requires no new primitive: stage is a path prefix (`/s0/.../s7/`), demotion is an ordinary `memories.update` that rewrites path and content, and because every update creates an immutable version, the pre-demotion text is preserved for free. Listing by `path_prefix` makes "everything at stage ≤ 3" a server-side query, and the store's `instructions` field steers read order.

3.2 ADMISSION: A DISJUNCTION, NOT A PRODUCT

Retrieval admits a memory when

$$\underbrace{\sigma \cdot \gamma^k \;\geq\; \theta}_{\text{default traversal}} \qquad \textbf{or} \qquad \underbrace{\sigma \;\geq\; \theta_{\text{strong}}}_{\text{cued override}}, \qquad \sigma = \exp\!\Big(-\tfrac{\|s_r - s^*\|^2}{2\delta_k^2}\Big),\; \delta_k = \delta_0(1+\eta k). \tag{2}$$

Here s^* is a contextual signal vector recorded at write time and δ_k widens with stage, so an aged trace is matched loosely by a gist-level cue.

The disjunction is essential and replaces a product form we originally specified. Under $\sigma \cdot \gamma^k \geq \theta$ alone, a stage-7 memory ($\gamma^k = 0.035$) requires $\sigma > 1.42$ to be admitted, which is impossible since $\sigma \leq 1$: old traces were unreachable at *any* cue strength, contradicting the model's own claim of recovery past τ. Repairing it moves overall accuracy on our corpus from 34.4% to 50.0% and flattens the profile across recent, mid-range and distant evidence from $100/33/25$ to $50/56/38$.

4 RECALL-AFTER-EVICTION

Why the obvious protocol fails. The natural test - sample questions, bucket by evidence age, compare under a budget - is uninformative, and we report this because it cost us several experiments. Distillation writes each dialogue turn into a median of **8** memories (mean 10.2, max 43); *no* question in our corpus depends on a single memory. Uniform random eviction of 65% of the store therefore destroys all copies of a question's evidence with probability $0.65^8 \approx 0.03$, and we measured exactly that: 8 of 382 questions. A random-eviction control matched the uncapped ceiling not because uniform thinning preserves recall, but because it had destroyed almost nothing.

The protocol. Condition on the capability. Build one long timeline, impose a byte budget B, and for each question determine whether the memories carrying its gold evidence survived. Evaluate only the questions whose evidence a destructive policy *erased*, asking every system the same set:

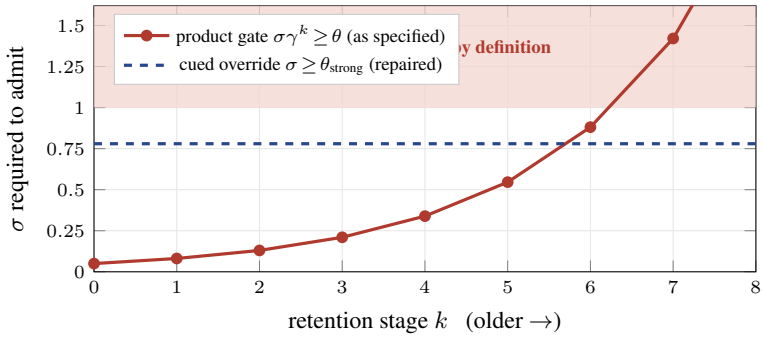


Figure 2: The admission rule forbade the recovery it was designed to provide. Under the product form, a stage-7 memory needs $\sigma > 1.42$ to be retrieved, which no cue can supply. The disjunction of Eq. (2) adds a flat override, so a sufficiently specific cue reaches an aged trace while weak cues still reach only recent ones.

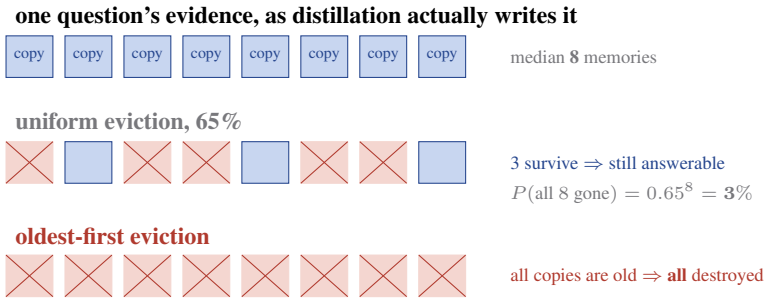


Figure 3: Why stratifying random questions by age measures nothing. Evidence is duplicated across a median of eight memories, so uniform eviction almost never destroys a question's support; recency-ordered eviction removes a contiguous block of time and takes every copy at once. A protocol must condition on evidence actually having been erased.

full	evidence intact - the ceiling
delete	evidence removed outright - nothing to retrieve
graded	evidence demoted and compressed - faint but present

The claim reduces to one falsifiable question: does a compressed gist retain enough to answer where a deleted memory retains nothing? A deletion policy cannot score here regardless of retrieval quality, which is the point: the protocol measures a capability rather than ranking retrievers.

5 Results

Setup. Three LoCoMo (Maharana et al., 2024) conversations concatenated into one 833-day timeline of 1,115 memories (135,243 bytes), budget $B = 35\%$. Backbone Claude-Sonnet-4.5; judge Claude-Haiku-4.5. Under *delete-oldest*, 248 of 382 questions lose their evidence entirely; their supporting turns have median age 675 days (range 200–833) against 74 days for the survivors, so the evicted set is precisely the distant past. We sample 40.

Against the strongest baseline. Mem0 (Chhikara et al., 2025) is the best-performing system we measured without a capacity bound, and it has no capacity policy: across 381 ingestions its reconciler issued 381 `ADD` operations and zero `UPDATE` or `DELETE`. Its store therefore grows linearly and overflows any budget, and placing it under one requires bolting on an eviction rule – we use its own `memories.delete`, oldest-first. Mem0 carries no dialogue identifiers, so the conditioned protocol of Sec. 4 cannot be applied to it; Table 2 instead holds both systems to the same absolute byte cap and stratifies questions by evidence age, which is the weaker design both representations admit.

Table 2: Mem0 versus graded forgetting under an identical 14,119-byte cap (conv-26, n=30). Mem0 must delete 279 of 381 memories to fit; graded forgetting deletes none. Accuracy by age of the oldest supporting turn.

System	recent 0–60d	mid 61–120d	old 121d+	All
Mem0, evicted to fit (102/381 kept)	60.0	6.2	25.0	26.7
Graded forgetting (321/321 kept)	**80.0**	**50.0**	**100.0**	**66.7**

Table 3: Recall of the evicted past. All three policies are asked the same questions; the two bounded policies are held to the same 46.2 KB footprint and differ only in how they reach it.

Policy	Store	Memories	Accuracy	Tok/q	How the budget is met
full (uncapped)	132.1 KB	1,115	75.0%	3,117	no bound – ceiling
delete-oldest	**46.2 KB**	397	**2.5%**	2,806	718 memories erased
graded forgetting	**46.2 KB**	**1,115**	**60.0%**	33,423	0 erased, detail shed

The ordering inverts. Without a bound Mem0 leads this corpus at 66.7% against 58.3% for flat retrieval over our store; held to the cap it reaches 26.7% while graded forgetting reaches 66.7%. Mem0's collapse is concentrated in the mid-range (6.2%), the signature of oldest-first eviction removing a contiguous block of the timeline. We note that this result is obtained under the protocol we criticise in Sec. 4 as insensitive; that it appears anyway, rather than being conditioned into existence, is the reason we report it.

Graded forgetting answers 60.0% of questions about material that deletion destroyed, recovering 80% of the uncapped ceiling from tier-2 gists while shedding the same 65% of bytes. Deletion's 2.5% is a single question answered from incidental surviving context. The gap is $24\times$ and is not a margin: no retrieval improvement can raise the deletion arm, because the evidence is gone.

The cost. Graded forgetting spends 33,423 tokens per query against 2,806. This is an implementation artefact - our reconstruction loop re-serialises the full accumulated evidence set each turn over a 1,115-memory graph - not a property of compression, but it is severe and unresolved, and we return to it in Sec. 6.

6 Limitations

A weak deletion baseline. *delete-oldest* is recency-ordered and unsophisticated. An importance-scored or LLM-judged policy would retain more of what matters, and the honest statement of our result is that compression beats *recency-ordered* deletion. We consider a learned eviction baseline the most important missing comparison.

One budget, one corpus, n=40. The effect is large (57.5 points) and will not vanish on resampling, but its shape across budgets is unmeasured; a sweep is the obvious next experiment.

Token cost. $12\times$ the deletion arm, from an implementation that re-sends accumulated evidence each turn. Deployment requires fixing this.

Confabulation. Reconstruction from a tier-ℓ memory is generative. We attach provenance tiers to every reconstructed span and treat τ as the boundary between *recalled* and *inferred*, but we have no calibration benchmark and regard one as a prerequisite for deployment.

Security. Retrieval-induced consolidation amplifies prompt injection: an injected memory that is retrieved is *promoted*, so content crafted to match frequent queries resists decay while legitimate memories sink past it. The platform documentation raises the underlying threat; our mechanism worsens it. Mitigations (consolidate only trusted-provenance writes; cap cumulative promotion; separate stores by trust level) are proposed but unevaluated.

7 RELATED WORK

Memory systems for LLM agents. Early persistent stores summarise dialogue into a vector index (Zhong et al., 2024; LangChain, 2025). Structured successors organise memory explicitly: A-Mem links LLM-authored notes into a graph (Xu et al., 2025); Zep maintains a bi-temporal knowledge graph that invalidates outdated edges (Rasmussen et al., 2025); AriGraph learns a knowledge-graph world model with episodic memory (Anokhin et al., 2024); MemoryOS imposes a short/mid/long-term hierarchy (Kang et al., 2025); SeCom segments memory at topically coherent boundaries (Pan et al., 2025); LiCoMemory uses a lightweight hierarchical index with temporal-aware search (Huang et al., 2025); and LightMem targets construction efficiency (Fang et al., 2025). Mem0 maintains a compact fact store through LLM-issued add/update/delete (Chhikara et al., 2025), and reflective memory management adds prospective and retrospective revision (Tan et al., 2025). All of these grow monotonically or prune destructively; none demotes. Recent surveys map the design space (Hu et al., 2025; Gao et al., 2026) and argue that long-term memory remains a principal deficiency of deployed agents (Hatalis et al., 2023; Hendrycks et al., 2025).

Retrieval, and retrieval as reconstruction. Retrieval-augmented generation established one-shot similarity retrieval over an external store (Lewis et al., 2020), later extended to graph-structured corpora (Han et al., 2025). Agentic variants move retrieval inside the reasoning loop, issuing queries on demand (Li et al., 2025) or learning to do so with reinforcement (Jin et al., 2025). Closest to our setting, Ji et al. (2026) recast memory access as active, multi-step reconstruction over a Cue–Tag–Content graph and prove active retrieval strictly more expressive than passive; we adopt that formulation for our reconstruction loop and treat its store as append-only, which is precisely the assumption we relax.

Deployed memory, and its retention policy. Anthropic's Memory Stores expose a versioned document store mounted into an agent sandbox, capped at 10,000 entries, with consolidation available as a separate *Dreams* job (Anthropic, 2026a;b). OpenAI's Agents SDK persists transcripts and compacts them when they grow (OpenAI, 2026a), while ChatGPT synthesises user memory in a background pass (OpenAI, 2026b). Each is a production system; each meets capacity by deleting, compacting or rewriting, and none offers graded, reversible demotion.

Benchmarks. LoCoMo provides long multi-session dialogues with annotated supporting evidence (Maharana et al., 2024); LongMemEval extends the horizon and adds knowledge-update questions (Wu et al., 2025). Both hold history at a fixed size. Recent analyses find full-context prompting remains a strong baseline at LoCoMo's scale (Wang et al., 2025; Zhang et al., 2025), which is consistent with our observation that capacity pressure - not question difficulty - is the regime where retention policy matters. Cross-paper comparison is further complicated by unstandardised judging protocols (Mem0, 2026), so we compare only within our own harness.

Cognitive foundations. The staged, cue-dependent view of forgetting descends from Ebbinghaus's retention curves (Ebbinghaus, 1885) and the distinction between storage and retrieval strength in the New Theory of Disuse (Bjork & Bjork, 1992), with ACT-R supplying an activation-decay formalism (Anderson & Lebiere, 1998). Encoding specificity holds that retrieval succeeds to the degree that cues reinstate the encoding context (Tulving, 1983; Godden & Baddeley, 1975), which motivates our signal vector. Reconstructive accounts from Bartlett (1932) to predictive coding (Clark, 2013) treat recall as assembly rather than readout; the neuroscience of retrieval describes cue-triggered reactivation of engrams (Frankland & Josselyn, 2019; Rugg & Renoult, 2025; Rashid et al., 2016), with episodic and semantic memory dissociable in the hippocampus (Manns et al., 2003) and recovery of apparently lost traces well documented (Aggleton & Brown, 1999; Proust, 1913). Content-addressable and sparse distributed memories established that stored patterns can be recovered from partial keys (Hopfield, 1982; Kanerva, 1988), which is the computational analogue of the cued override in Eq. (2).

8 CONCLUSION

Deployed agent memory meets capacity by deleting, irreversibly, without knowledge of future queries. We showed that it need not: under a hard budget, shedding detail rather than items preserves 60.0%

answerability on material that deletion reduces to 2.5%, at identical storage cost. We also showed that Mem0, the strongest published baseline, has no capacity policy of its own and must meet a budget by deleting, which costs it two thirds of its accuracy.

We take the combination seriously rather than apologetically. It says the value of graded forgetting is not in answering better but in *still being able to answer at all* once the store will not fit - and that benchmarks which hold history at a fixed, comfortable size cannot see the difference. The costs are paid on every query; the benefit appears only at capacity. Evaluating agent memory on longer histories, rather than harder questions, is in our view the change the field most needs.

Reproducibility statement

The software artifacts of this paper are available at `https://anonymous.4open.science/r/agent-memory-2D21/README.md`.

Use of AI in this work

Per the ICLR 2027 AI Policy for Authors, we disclose the following. A large language model was used substantially throughout this project, in categories the policy lists as requiring disclosure.

Methodology and experiment design. The model proposed the recall-after-eviction protocol, identified that age-stratified eviction is uninformative under evidence redundancy, and designed the eviction policies and their controls.

Implementation. All experimental code - the memory lifecycle, the retrieval agents, the benchmark harnesses, and the analysis scripts - was written by the model under author direction.

Interpretation of results. The model performed the analyses reported in Sections 5 and 6, including the diagnosis of the admission-gate defect, the evidence-redundancy measurement, and the date-encoding finding.

Writing. The model drafted this manuscript, including its figures.

Experimental subject. Claude Sonnet 4.5 and GPT-5.6 were additionally used as the agent backbones under evaluation, and Claude Haiku 4.5 as the LLM judge; these are experimental apparatus rather than authoring assistance.

The authors take full responsibility for all content, including every claim, number and citation. All quantitative results were produced by executing the released code; all citations were verified to resolve to real publications.

References

J. P. Aggleton and M. W. Brown. Episodic memory, amnesia, and the hippocampal–anterior thalamic axis. *Behavioral and Brain Sciences*, 22(3):425–444, 1999.

J. R. Anderson and C. Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, 1998.

P. Anokhin, N. Semenov, A. Sorokin, D. Evseev, A. Kravchenko, M. Burtsev, and E. Burnaev. AriGraph: Learning knowledge graph world models with episodic memory for LLM agents. *arXiv:2407.04363*, 2024.

Anthropic. Using agent memory - Managed Agents documentation, 2026. `https://platform.claude.com/docs/en/managed-agents/memory`.

Anthropic. Dreams: memory consolidation for Managed Agents, 2026. `https://platform.claude.com/docs/en/managed-agents/dreams`.

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

R. A. Bjork and E. L. Bjork. A new theory of disuse and an old theory of stimulus fluctuation. In *From Learning Processes to Cognitive Processes*, volume 2, pp. 35–67. Lawrence Erlbaum, 1992.

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.

A. Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

H. Ebbinghaus. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot, 1885.

J. Fang, X. Deng, H. Xu, Z. Jiang, Y. Tang, Z. Xu, S. Deng, Y. Yao, M. Wang, S. Qiao, H. Chen, and N. Zhang. LightMem: Lightweight and efficient memory-augmented generation. *arXiv:2510.18866*, 2025.

P. W. Frankland and S. A. Josselyn. The neurobiological foundation of memory retrieval. *Nature Neuroscience*, 22(10):1576–1585, 2019.

H. Gao et al. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *Transactions on Machine Learning Research*, 2026.

D. R. Godden and A. D. Baddeley. Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, 66(3):325–331, 1975.

H. Han et al. Retrieval-augmented generation with graphs (GraphRAG). *arXiv:2501.00309*, 2025.

K. Hatalis, D. Christou, J. Myers, S. Jones, K. Lambert, A. Amos-Binks, Z. Dannenhauer, and D. Dannenhauer. Memory matters: The need to improve long-term memory in LLM-agents. In *Proceedings of the AAAI Symposium Series*, volume 2, pp. 277–280, 2023.

D. Hendrycks et al. A definition of AGI. *arXiv:2510.18212*, 2025.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

Y. Hu, S. Liu, Y. Yue, G. Zhang, B. Liu, F. Zhu, J. Lin, H. Guo, S. Dou, and Z. Xi. Memory in the age of AI agents. *arXiv:2512.13564*, 2025.

Z. Huang, Z. Tian, Q. Guo, F. Zhang, Y. Zhou, D. Jiang, and X. Zhou. LiCoMemory: Lightweight and cognitive agentic memory for efficient long-term reasoning. *arXiv:2511.01448*, 2025.

S. Ji, Y. Li, and B. Hooi. Memory is reconstructed, not retrieved: Graph memory for LLM agents. In *Proceedings of the 43rd International Conference on Machine Learning*, PMLR 306, 2026.

B. Jin, H. Zeng, Z. Yue, D. Wang, H. Zamani, and J. Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv:2503.09516*, 2025.

P. Kanerva. *Sparse Distributed Memory*. MIT Press, 1988.

J. Kang, M. Ji, Z. Zhao, and T. Bai. Memory OS of AI agent. *arXiv:2506.06326*, 2025.

LangChain. LangMem SDK for agent long-term memory, 2025. `https://blog.langchain.com/langmem-sdk-launch/`.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, 2020.

X. Li, G. Dong, J. Jin, Y. Zhang, Y. Zhou, Y. Zhu, P. Zhang, and Z. Dou. Search-o1: Agentic search-enhanced large reasoning models. In *EMNLP*, pp. 5420–5438, 2025.

A. Maharana, D. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang. Evaluating very long-term conversational memory of LLM agents. In *ACL*, pp. 13851–13870, 2024.

J. R. Manns, R. O. Hopkins, and L. R. Squire. Semantic memory and the human hippocampus. *Neuron*, 38(1):127–133, 2003.

Mem0. LoCoMo vs. LongMemEval vs. BEAM: the 2026 AI memory benchmark guide, 2026. `https://mem0.ai/blog/ai-memory-benchmarks-in-2026`.

OpenAI. Sessions - OpenAI Agents SDK documentation, 2026. `https://openai.github.io/openai-agents-python/sessions/`.

OpenAI. Dreaming: better memory for a more helpful ChatGPT, 2026. `https://openai.com/index/chatgpt-memory-dreaming/`.

Z. Pan, Q. Wu, H. Jiang, X. Luo, H. Cheng, D. Li, Y. Yang, C.-Y. Lin, H. V. Zhao, and L. Qiu. SeCom: On memory construction and retrieval for personalized conversational agents. In *ICLR*, 2025.

M. Proust. *Du côté de chez Swann*. Grasset, 1913.

A. J. Rashid, C. Yan, V. Mercaldo, H.-L. Hsiang, S. Park, C. J. Cole, A. De Cristofaro, J. Yu, C. Ramakrishnan, and S. Y. Lee. Competition between engrams influences fear memory formation and recall. *Science*, 353(6297):383–387, 2016.

P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, and D. Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv:2501.13956*, 2025.

M. D. Rugg and L. Renoult. The cognitive neuroscience of memory representation. *Neuroscience & Biobehavioral Reviews*, 155:105450, 2025.

Z. Tan, J. Yan, I.-H. Hsu, R. Han, Z. Wang, L. Le, Y. Song, Y. Chen, H. Palangi, and G. Lee. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *ACL*, pp. 8416–8439, 2025.

E. Tulving. *Elements of Episodic Memory*. Oxford University Press, 1983.

A simple yet strong baseline for long-term conversational memory of LLM agents. *arXiv:2511.17208*, 2025.

D. Wu, H. Wang, W. Yu, Y. Zhang, K. Chang, and D. Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *ICLR*, 2025.

W. Xu, Z. Liang, K. Mei, H. Gao, J. Tan, and Y. Zhang. A-MEM: Agentic memory for LLM agents. *arXiv:2502.12110*, 2025.

Evaluating long-term memory for long-context question answering. *arXiv:2510.23730*, 2025.

W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang. MemoryBank: Enhancing large language models with long-term memory. In *AAAI*, pp. 19724–19731, 2024.