Reading Committed State When Memory Is Provenanced

A Fathom coherence read run against the reading discipline of Agent Zero Memory, on LongMemEval.

Embedded Risk Analytics · A Fathom Case Study · September 2026

The coherence of a long-horizon agent is a measurable property of the state it commits. A system that acts on a record it has itself produced carries a reflexive burden, an informational cost that grows as the system's own actions shape what it must later model (Galligan, 2026). When that burden grows faster than the capacity of the record layer that holds it, the committed state stops constraining the agent's next step, and coherence fails.

Long-term memory for agents is advancing quickly, and Agent Zero Memory (Zhu and Wu, 2026) is a strong recent example. It keeps an episodic timeline, an entity-event graph, and a curated documentary memory side by side over the same history, gives every stored item its origin, timestamp, and evidence pointer, and holds answers under a citation lock: a reply may cite only evidence the reader actually opened, and a reader that cannot assemble such citations abstains. That is a serious discipline, and it addresses grounding well. This study measures the property it leaves open, whether a grounded answer is also current with the record. We ran that measurement against the paper's own reading rules, on real data, across two models.

The expectations based on theory

Our framework states a feasibility condition for any system that must maintain a coherent history:

$$C \geq H_{ext} + B_{\rho}$$

Here C is the effective capacity of the record layer, H_{ext} is the entropy the environment imposes, and B_{ρ} is the reflexive burden, the cost of tracking the system's own causal influence on what it will later read (Galligan, 2026). Provenance raises the effective capacity of the record: every item is findable, dated, and traceable to its source, and nothing is overwritten. A citation lock then guarantees that what the reader says is supported by something in that record.

Our framework predicts that this leaves the reflexive-burden term unaddressed. In a non-destructive timeline the superseded value of a fact is a fully provenanced, fully citable item, and the abstain rule fires only when evidence is missing, never when it is stale. So a reader can hold a valid citation lock and still commit a value the record has already replaced. LongMemEval's knowledge-update questions (Wu

et al., ICLR 2025) put that prediction on real data, since each one states a value and then revises it, and the correct answer is the revised value.

What we tested

No code accompanies the paper, so we implemented the reading discipline as the paper states it and nothing beyond it: an append-only timeline in which every item carries an origin, an immutable timestamp, and an evidence pointer; hybrid lexical retrieval (BM25 plus fuzzy match, fused by reciprocal rank at k=60) with a relevance rerank; and a reader that sees the top-k items, may open more items once, and must either answer with citations drawn from what it opened or abstain. Per episode we appended every turn of every session for three real knowledge-update facts to one timeline, roughly 70 items, and asked the three questions at k of 2, 4, 8, and the whole timeline.

For every answer we measured three quantities on the paper's own terms. The first is the citation lock, verified mechanically: the cited items are a subset of the opened items, and a cited item carries the value the reader answered. The second is task accuracy, whether the answer matched the updated value. The third is the committed-state read over the full timeline, which flags an answer that references a value the record has superseded, or whose cited items all predate the event that established the current value. The read never sees the benchmark's gold answer; it folds the record's own distilled events. We ran the study on two models of different lineage, deepseek-chat and gpt-4o-mini, through OpenRouter.

A citation-locked answer goes stale when the update falls outside the window

The abstain rule did what the paper says it does. At k=2 most questions came back as abstentions, because the two items shown carried no value at all. The failure appears in the other case, where the window held the original statement and not the update. There the reader almost never abstained. It cited the superseded item and answered with the superseded value, and every one of those answers passed the citation lock on verification.

Model	k	Answered	Abstained	Correct	Stale	Stale and lock-verified
deepseek-chat	2	15 / 44	29 / 44	6	9	9 / 9
	4	30 / 44	14 / 44	17	13	13 / 13
	8	40 / 44	4 / 44	25	14	14 / 14
	all	44 / 44	0 / 44	38	6	6 / 6
gpt-4o-mini	2	14 / 44	30 / 44	6	8	8 / 8
	4	29 / 44	15 / 44	17	12	12 / 12
	8	38 / 44	6 / 44	25	13	13 / 13

Model	k	Answered	Abstained	Correct	Stale	Stale and lock-verified
	all	43 / 44	1 / 44	35	8	8 / 8

Table 1. Knowledge-update questions from LongMemEval (oracle subset), 44 facts in 15 episodes of three, per model, under the paper's retrieval turn at each window size k. Counts are questions. "Stale" means the answer matched the value the record had superseded; "lock-verified" means the cited items were among those opened and a cited item carried the answered value.

Restricting to the questions where the update item lay outside the retrieval window and the reader chose to answer, the stale rate was 9 of 9, 13 of 13, and 11 of 13 for deepseek-chat at k of 2, 4, and 8, and 8 of 8, 11 of 12, and 9 of 11 for gpt-4o-mini. Superseded evidence is evidence, so the lock holds and the abstain rule stays quiet. In our tests the lock never once rejected a stale answer, on 83 stale answers across the two models.

Relevance and currency pull in opposite directions

With the entire timeline visible, the reader still returned the superseded value on 6 of 44 questions for deepseek-chat and 8 of 44 for gpt-4o-mini, each of them lock-verified. Reading those cases by hand shows the pattern. The original statement tends to be on topic, a user asking for gym reminders at 7:00 pm, or for closing-cost estimates on a $350,000 pre-approval. The update tends to arrive as a passing mention inside an unrelated session, a meeting that has to end before the gym at 6:00 pm, a note that the pre-approval came back at $400,000. A relevance rerank ranks the original higher because it matches the question better, and the paper skips the rerank only for explicitly temporal queries. The knowledge-update questions are not phrased that way. The design that makes the memory findable is the same design that surfaces the stale item first.

A currency check is one clause away, and it is not quite enough

The generous reading of this result is that the paper already stores what a currency check needs. Every cited item has a timestamp, and the timeline knows when the latest event on a thread occurred. A rule that flags any answer whose cited items all predate that event would have caught 41 of the 42 stale answers for deepseek-chat and 38 of the 41 for gpt-4o-mini, with one firing on an answer that was neither the old nor the new value but cited only pre-update evidence, which is arguably a correct flag.

Model	Stale answers	Caught by cited-item timestamps	Cited both old and new, committed the old value	Fired on non-stale answers
deepseek-chat	42	41	1	1
gpt-4o-mini	41	38	3	0

Table 2. The timestamp-only currency check across all four window sizes. Misses are answers that cited the update item alongside the original and still committed the original value.

The misses are the instructive part. In those cases the reader opened both the original and the update, cited both, and committed the old value anyway. Provenance cannot see that, because the citation set is current. Only a read of the value the agent actually committed, against the value the record currently holds, catches it. That is the committed-state read, and it fired on every stale answer in the study with no firings on coherent answers.

A single-probe score misses stale commitments elsewhere in the record

One question grades one fact. Because each episode held three facts on one timeline, we could also ask whether a graded question can pass while another commitment in the same record is stale. Taking the second fact of each episode as the probe, the probe answer was correct while the read fired on a different fact in the episode on 2, 3, 4, and 4 of 15 episodes for deepseek-chat at k of 2, 4, 8, and all, and on 2, 4, 4, and 2 of 15 for gpt-4o-mini. A benchmark score of 95.60 percent is a single-probe score, and it has no way to report this cell.

Where our Fathom instrument sits

Provenance and the citation lock act on capacity. They make the record larger, better indexed, and traceable, and they guarantee that an answer is grounded in it. The reflexive burden is a separate term, and our framework identifies it as the one that binds for long-horizon agents. Fathom measures it. It sits beside the memory layer, reads the record the agent commits, and reports where that record cites or commits a value the agent has already superseded. Holding several commitments consistent as the retrieval window moves is the multi-fact form of reflexive burden, and the read separates that burden from the accuracy of any single answer.

Boundary

These results come from our implementation of the reading discipline as the paper describes it. The paper releases no code, prompts, or per-category results, and it does not name the LongMemEval variant or the judge it used, so the numbers here describe our harness and not the authors' system. Our retrieval is lexical only; the paper also uses an embedding channel, and a memory that distills and links events may retrieve the update more often than raw turns do. The samples are small and cover two models on one benchmark. Three corrections were applied after the runs and are disclosed in the harness: a window flag was recomputed with value-distinct token matching, one fact was excluded because its extracted pair was not a value change, and one extracted prior value was corrected from the record. The committed-state read reported here is the shipping read. The information-theoretic decomposition and its scoring stay behind the hosted instrument.

Our position

Provenance-aware memory is a real advance, and a citation lock is the right discipline for grounding. It also raises the value of a coherent committed state, since a system that can prove every answer is supported has more room to be confidently wrong about which support is current. Fathom is the read that tells you whether it was. We run it on live third-party runtimes without an oracle. If you build long-horizon or memory-backed agents and want to know whether the state they commit stays coherent with the record they keep, we would be glad to run a committed-state read on a sample of your traces and show you where the record holds and where it drifts.

References and notes

Peter Galligan. "Records, Reflexive Modeling, and the Nomological Conditions for Stable Physical Histories." SSRN working paper 6683578, 2026. The feasibility inequality and the reflexive-burden construct used to motivate this study are drawn from that paper.

Pengyuan Zhu, Ming Wu. "Agent Zero Memory: Provenance-Aware Long-Term Memory for LLM Agents." arXiv:2608.29606, 2026. Zero Labs. No code or data release accompanies the paper; the reading discipline used here is implemented from the paper's Definitions 1 and 2 and its description of retrieval, with the implementation in our harness.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, Dong Yu. "LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory." ICLR 2025. arXiv:2410.10813. Data used under the MIT license.

Models accessed through OpenRouter: deepseek-chat (DeepSeek) and gpt-4o-mini (OpenAI). Figures are from our runs on the dates of this study and describe those runs only.