Reading Committed State When an Agent Manages Its Own Context

A Fathom coherence read run against a ContextPilot-style context-management step, 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.

A long-horizon agent is exactly this kind of system. Proactive context management is advancing quickly, and ContextPilot (Pan et al., EMNLP 2026) is a strong recent example, training an agent to edit its own working context and reporting higher task accuracy on a more compact context. That work manages the capacity side of the constraint. This study measures the side it leaves open, whether the state the agent commits stays coherent with itself. In this case study, we run that measurement against ContextPilot's own operation, 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). In low-interaction regimes, the reflexive burden is small and the condition reduces to a familiar bound on record capacity. In the regime a long-horizon agent occupies, where its own commitments increasingly become the context it acts on, the reflexive burden grows and becomes the binding term. Context management raises effective capacity and lowers the entropy the agent carries forward.

Our framework predicts that this leaves the reflexive-burden term unaddressed, so an agent can hold a compact, well-managed context and still commit a state that is inconsistent with what it has already decided. LongMemEval's knowledge-update questions (Wu et al., ICLR 2025) put that prediction on real data, since each one

commits the agent to a value and then revises it, and the correct answer is the revised value.

What we tested

We applied ContextPilot's foldHistory operation, which discards prior history and builds a searchable index, to the session that carried the update, and we left the earlier session in view. The agent could search the folded index to recover the update before it answered. For every run we measured two quantities on their own terms. The first is task accuracy, whether the answer matched the updated value. The second is the committed-state read, whether the answer the agent committed referenced a value the record had already superseded. We ran the study on two models of different lineage, deepseek-chat and gpt-4o-mini, through OpenRouter.

Folding the update turns a correct answer into a confident stale one

With the full history in context, both models answered the updated value on nearly every question. After foldHistory, accuracy fell sharply on both, and most of the failures were confident reversions to the superseded value. Acknowledgments of uncertainty were rare, and searches to recover the update from the folded index were rarer still.

Model	Correct, full history	Correct, after foldHistory	Reverted to superseded value	Searched to recover
deepseek-chat	39 / 40	16 / 40	22 / 40	4 / 40
gpt-4o-mini	40 / 40	14 / 40	24 / 40	1 / 40

Table 1. Knowledge-update questions from LongMemEval (oracle subset), 40 per model, with foldHistory applied to the update session and search available. Counts are questions.

In our tests the pattern held across both models. Roughly 55 to 60 percent of the questions came back with the value the agent had already replaced, and the models retrieved the update they needed on fewer than one in ten runs even though the folded index stayed searchable throughout. The capacity-side operation worked as intended and the committed answer went stale anyway, which is the behavior the reflexive-burden term predicts.

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

One question grades one fact. To see whether the committed record stays coherent beyond the fact under test, we composed three knowledge-update facts into one context, folded one of them, and graded a probe question on a different fact whose update stayed visible. The probe answer stands in for a single-probe evaluation. The committed-state read audits all three committed answers. The additive case is

a run where the probe answer is correct and the read still fires, because a fact the probe never touched was left stale.

Model	Probe correct and read fires (additive)	Folded fact left stale	Read false positives, verified
deepseek-chat	6 / 10	9 / 10	0
gpt-4o-mini	7 / 10	9 / 10	0

Table 2. Composite runs, 10 episodes of three facts per model. The victim fact's update is folded; the probe is graded on a different, visible fact. Read false positives count firings on coherent committed state, checked case by case.

The read stayed silent whenever the committed state was coherent. Across more than one hundred runs on the two models it produced no false positives on inspection. Its firings in the condition where nothing was folded corresponded to genuine reversions the models made under the load of holding several facts at once, and we confirmed each of those by hand. Loading three facts into one context and asking about them was enough to produce stale commitments on its own, on 2 of 10 episodes for deepseek-chat and 5 of 10 for gpt-4o-mini, which the read caught while the graded question passed. The burden shows up from interaction density alone, and it rises when the context is folded.

Where our Fathom instrument sits

foldHistory (and operations like it) act on capacity and on the environmental entropy the agent carries forward. They make the record smaller and the carried entropy lower, and they do that well. 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 context-management layer, reads the record the agent commits, and reports where that record cites a value the agent has already superseded. Holding several commitments consistent as the context is edited is the multi-agent, multi-fact form of reflexive burden, and the read separates that burden from the accuracy of any single answer.

Boundary

On a single-answer question, a reversion to the superseded value is also a wrong answer, so on one isolated fact the read and the accuracy metric agree. The read earns its value on the committed record as a whole, where a graded question can pass while another commitment is left stale, which is the case Table 2 measures. These results apply foldHistory in a controlled harness built on ContextPilot's own operation, and the trained ContextPilot agent may manage its context differently. The samples are small and cover two models on one benchmark. The committed-state read reported here is the shipping read. The information-theoretic decomposition and its scoring stay behind the hosted instrument.

Our position

Context management is a real advance, and the tools shipping for memory and compression make long-horizon agents cheaper and more capable. They also raise the value of a coherent committed state, since an agent that edits its own context aggressively has more room to commit to a value it has already replaced. Fathom is the read that tells you whether it did. We run it on live third-party runtimes without an oracle. If you build long-horizon or multi-agent systems and want to know whether the state they commit stays coherent, 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.

Zhuoshi Pan, Qizhi Pei, Junru Lu, Honglin Lin, H. Vicky Zhao, Di Yin, Xing Sun. “ContextPilot: Teaching Agents for Proactive Context Management via Fine-grained RL.” arXiv:2608.28476, 2026. Accepted to EMNLP 2026 (Main Track). Code released under Apache-2.0. The foldHistory operation used here is adapted from that release, with attribution and changes noted 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.