DEEP-DREAM: A TEMPORAL CONCEPT GRAPH FOR AGENTIC LONG-TERM MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

An agent's long-term memory grows while the entities it mentions may change substantially. Human recall does not rewrite an earlier observation when later knowledge arrives: it retrieves that observation in the historical context in which it was made. An agent faces the same problem under a finite context window. Copying the whole history is too expensive, while summarizing it before the question is known can discard the qualifiers that make an answer correct. Deep-Dream is designed around these two requirements. It leaves source documents and spans unchanged, and maintains concepts, relations, observations, and versions as a linked access layer. At query time, the agent follows these links across documents, opens the source spans that are needed, and can continue to a multi-hop trail when the first evidence is insufficient. During ingestion, uncertain mentions are kept as candidate entity families instead of being forced into one identity; later observations and relation neighborhoods can align or redirect those candidates without deleting the original evidence. We evaluate this design on MemoryAgentBench, LongMemEval, LoCoMo, and MEME. Source-grounded reading reaches 71.0% on MemoryAgentBench, close to the 78.7% same-model full-context reference, and Deep-Dream reaches 95.19% on LoCoMo. Delayed entity alignment reduces ingest calls by 46–65% while retaining comparable sampled quality. The results show how a revisable access layer and agent-controlled source reading support long-term memory; MEME also shows that deletion requires an explicit suppression mechanism.

1 INTRODUCTION

A memory answer has at least three evidential layers: what a document says, how the system links that sentence to an entity, and which passages the agent actually reads. Conflating them makes a plausible answer hard to audit. Suppose a document says, "Alex approved the contract,"and the collection contains an Alex at a laboratory and another at a vendor. The sentence supports an approval claim, but not an identity assignment. Later organization or relation evidence may change the link. A premature merge can make the wrong history appear authoritative.

The same calibration problem occurs with updates. A friend says in January, "I am avoiding spicy food for now,"and in June, "I can eat spicy food again."Both statements are evidence. A current answer should use the later one, while an explanation of the change should inspect both. Retention and interpretation therefore need separate state.

Deep-Dream implements this separation as a *document-first temporal concept graph*. Documents, Episodes, and exact source offsets preserve the record. Observations link extracted content to spans; concepts, relations, versions, and candidate families provide revisable access paths. New evidence can redirect a link without rewriting its source. Processing time orders updates, and dates remain observation content. This is an auditable representation, not a guarantee of correct identity resolution or historical-state recovery.

At query time, the agent follows an address, opens its source passage, and decides whether the evidence is sufficient (Figure 1). It may search again, inspect a relation neighborhood, or compare earlier and later wording. **The source gate validates provenance, not identity or entailment.** This distinction keeps a source citation from being mistaken for a proof of the answer.

We evaluate both axes with paired construction runs, MemoryAgentBench read paths, LongMemEval and LoCoMo history tests, and MEME deletion checks. The comparison reports what each protocol establishes: complete paths are not component ablations, duplicate-name counts are proxies, and provenance is not entailment. Our contributions are:

- **Calibrated evidence storage.** Source-linked observations remain inspectable while inferred identity links can be revised.

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.

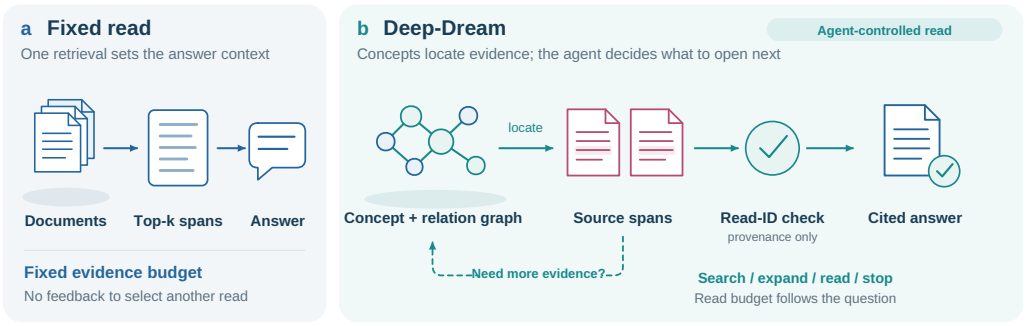

Figure 1: Storage and reading expose different evidence claims. Revisable addresses guide the search; source passages let the agent audit wording, identity context, and update status.

- **Auditable progressive reading.** The agent can gather additional passages and expose the evidence behind an address.
- **A two-axis evaluation.** We report quality, coverage, costs, alignment uncertainty, null graph recall, and deletion failure with their protocol boundaries.

2 Related Work

Long-term memory benchmarks. LoCoMo and LongMemEval evaluate long conversations, temporal updates, and multi-session retrieval (Maharana et al., 2024; Wu et al., 2024). MEME adds questions about deletion, cascade effects, and absent facts (Jung et al., 2026). MemoryAgentBench (MAB) combines accurate retrieval, temporal and test-time learning, long-range understanding, and selective forgetting in one evaluation framework (Hu et al., 2026). These benchmarks measure different parts of the storage and reading problem, so we use them as complementary tests rather than reducing memory to one score.

Memory systems for agents. MemGPT treats context management as paging between a working window and an external store (Packer et al., 2023). Model-Native Computing Architecture likewise frames context capacity, cache reuse, and agent scheduling as system-level resource-management problems (Lin et al., 2026). Mem0, Zep/Graphiti, HippoRAG-2, A-MEM, MIRIX, and Cognee use summaries, notes, graphs, or combinations of them to persist information across sessions (Chhikara et al., 2025; Rasmussen et al., 2025; Gutiérrez et al., 2025; Xu et al., 2026; Wang & Chen, 2025; Markovic et al., 2025). These systems differ in what they keep as the memory object: Deep-Dream keeps the source document and treats concepts and relations as links back to it, so a derived graph entry is a search handle rather than a replacement fact.

Temporal and versioned knowledge. Temporal knowledge graphs represent changing entities and time-dependent relations, while probabilistic record linkage separates false merges from missed matches (Fellegi & Sunter, 1969). Deep-Dream applies these ideas to agent memory: processing time orders observations and versions, and delayed alignment allows a candidate identity to remain revisable while its source evidence stays intact. Event dates remain part of the observations and can guide retrieval without creating a second state clock.

Compression, retrieval, and evidence. Information bottleneck and rate–distortion theory describe the trade between a compact representation and information needed for a target (Tishby et al., 2000; Shannon, 1993). Associative-memory models study retrieval from partial cues (Ramsauer et al., 2020), while retrieval-augmented generation selects external evidence for generation (Lewis et al., 2020). Long-context models can lose relevant information when it is buried in a large window (Liu et al., 2024); attribution work checks whether a response cites a source (Gao et al., 2023). Deep-Dream combines these ideas in a persistent setting: concepts and relations narrow the search, source spans retain the wording, and the agent chooses how much evidence to load; Section 3.5 formalizes this separation.

Agentic retrieval. A-RAG exposes keyword, semantic, and chunk-reading tools to a reasoning model for multi-hop retrieval (Du et al., 2026). Deep-Dream shares the idea that a model can choose its next retrieval action, but applies it to a changing personal history. Its additional requirements are identity revision, version links, cross-document episodes, and source-grounded verification.

3 DEEP-DREAM

3.1 MEMORY REPRESENTATION

Let $\mathcal{D}_t$ be the documents processed by time t. For a document $D \in \mathcal{D}_t$, let $\mathcal{E}(D)$ be the Episodes obtained from its immutable source spans. For an Episode $e_i \in \mathcal{E}(D_i)$, define one stored observation as

$$o_i = (D_i, e_i, \tau_i, x_i, \ell_i),$$

where τ_i is processing time, x_i is the extracted content and attributes, and ℓ_i is the exact source offset in D_i. The observation store after t writes is $O_t = \{o_1, \ldots, o_t\}$. The persistent state is

$$M_t = (F_t, O_t, G_t),$$

where F_t contains concept families and candidate identities, and G_t contains relation, neighborhood, and version pointers. Every observation carries the document and Episode identifiers (D_i, e_i) that recover its source text.

A write step is represented by

$$\begin{aligned} F_{t+1} &= \mathsf{Align}(F_t, x_{t+1}), \\ O_{t+1} &= O_t \cup \{o_{t+1}\}, \\ G_{t+1} &= \mathsf{Link}(G_t, F_{t+1}, o_{t+1}), \\ M_{t+1} &= (F_{t+1}, O_{t+1}, G_{t+1}). \end{aligned} \tag{1}$$

Here Align either assigns x_{t+1} to an existing family or keeps a new candidate family; Link adds or redirects pointers between families, observations, and versions. Neither operation edits the text addressed by ℓ_{t+1}. Under the representation invariant that version pointers retain the predecessor state, a projection π_t can recover M_t from M_{t+1}. This is a conditional property of the retained state, not an empirical guarantee that the implementation can recover every historical graph configuration; updates add links without deleting the earlier observation record.

This representation separates finding from reading: resolving a family or relation is an index lookup, while opening its source span is a separate load. A relation edge helps the agent find a connected observation, but the assertion supporting it remains in the source text. Processing time τ_i is the single state axis of the graph; an event date may appear in x_i and help answer a question, but it does not create a second version clock.

3.2 WRITING MEMORY

Writing follows the state transition in Equation (1): documents are split into Episodes with source offsets; entities and relations are extracted from each Episode; candidate identities are compared with existing concept families; and accepted links are written together with their supporting observations. Every link therefore has a path back to a document span.

When the first mention of an entity is ambiguous, Deep-Dream keeps candidate families separate. Later observations and relation neighborhoods can provide enough evidence to merge compatible families or redirect their pointers. The operation changes F_t and G_t, not the observations in O_t. A temporary split is therefore recoverable, whereas an incorrect early merge can mix two histories under one search handle.

3.3 IMMEDIATE AND DELAYED ENTITY ALIGNMENT

The experiments compare two write schedules over the same state representation. **Immediate alignment (Base)** applies Align as each window arrives. **Delayed alignment (Align+)** first appends o_i and candidate links, then resolves compatible families after the surrounding windows have been processed. The difference is the time of the identity decision, not the schema or source record.

Delayed alignment uses more context before making a merge. In our implementation it is bundled with batched or concurrent processing, so the paired measurements in Section 4.2 quantify the deployed bundle rather than delayed alignment in isolation. It can reduce repeated work while preserving the source spans that support each decision.

3.4 READING MEMORY

Let $\mathcal{S}_t$ be the set of immutable source spans in M_t. For a query q, the agent starts with $C_0(q) = \varnothing$ and chooses an action

$$a_{k+1} = \rho\big(q, M_t, C_k(q), u_k\big) \in \big(\mathcal{S}_t \setminus C_k(q)\big) \cup \{\mathsf{stop}\},$$

where u_k is the agent's current uncertainty. If a_{k+1} is a span, it is opened and

$$C_{k+1}(q) = C_k(q) \cup \{a_{k+1}\};$$

if $a_{k+1} = \mathsf{stop}$, the chain ends. By induction—each non-stop step adds one member of $\mathcal{S}_t$ to C_k—we have $C_k(q) \subseteq \mathcal{S}_t$ and $C_k(q) \subseteq C_{k+1}(q)$. The equation tracks source-opening actions; lexical, semantic, concept, and relation lookups update u_k without adding a span to C_k. This formalizes the source-opening trace; the policy decides when the evidence is sufficient. The information terms below are identities from the chain rule, not estimates of navigation utility or answer quality.

At query time, the agent begins with the cheapest useful cue—lexical, semantic, concept, or relation—opens the corresponding Episode or source span, and expands the search when the evidence is incomplete: a question about one fact may stop after one passage, while an update or multi-document question may follow several relation and version links.

We evaluate three read paths:

1. **Fixed retrieval** runs the lexical/semantic/graph retriever once and gives the answerer one ranked list of memory records and retrieved chunks. The list cannot be changed or checked against a source file.
2. **Memory-tool reading** gives the answering agent read-only search, concept, relation, and chunk tools. It may call them repeatedly and use one result to form the next query, but the answerer receives the records returned by those tools.
3. **Source-grounded reading** uses the index to select document groups, opens the original files, reads exact spans, and submits only spans that pass the source gate.

The source check verifies that a cited identifier was returned by a tool and still points to the original text. It establishes provenance of the citation; the agent must still decide whether the passage supports the answer.

3.5 Why this representation preserves useful detail

Let Z be a fixed target assignment of observations to entities. The dimension of Z is held fixed as observations arrive. Since $\pi_t(M_{t+1}) = M_t$, the newer state contains the previous state together with the new observation and any revised links. Therefore

$$\begin{aligned} H(Z \mid M_t) - H(Z \mid M_{t+1}) &= I(Z; M_{t+1} \mid M_t) \\ &\geq 0. \end{aligned} \tag{2}$$

The first equality holds because M_t is recoverable from M_{t+1}—a representation invariant supplied by the retained version links, not by arbitrary graph edges. If the new observation carries no information about Z, the difference is zero; when it distinguishes candidate families, the conditional entropy decreases. This is a structural guarantee about retained state, not a measured entropy gain.

For a query q, let A be the compact record used to locate candidates, E_C the source spans opened from candidate set C, and Y the answer-relevant fact. The chain rule gives

$$I(Y; A, E_C \mid Q = q) = I(Y; A \mid Q = q) + I(Y; E_C \mid A, Q = q). \tag{3}$$

This is the chain rule. The first term is information used to locate a candidate; the second is the additional information obtained by reading its source. Deep-Dream compresses the search description, not the only copy of the fact.

Let R_q be the event that a relevant source span is in the opened set $C_K(q)$, and let V_q be the event that the produced answer is correct, $V_q = \{\widehat{Y} = Y\}$. By the law of total probability,

$$\begin{aligned} \Pr[\widehat{Y} = Y \mid q] &= \Pr[R_q \mid q] \Pr[V_q \mid R_q, q] \\ &\quad + \Pr[\neg R_q \mid q] \Pr[V_q \mid \neg R_q, q] \\ &\geq \Pr[R_q \mid q] \Pr[V_q \mid R_q, q]. \end{aligned} \tag{4}$$

Concept and relation links primarily affect the first factor; source reading and citation checking affect the second. The inequality is not a performance estimate or a claim that retrieval is necessary for a correct answer; it only decomposes outcomes according to whether a relevant span was opened.

4 Experiments

We evaluate the two operations separately: storing a growing history with updates and revisable identities, and reading only the evidence required by a question.

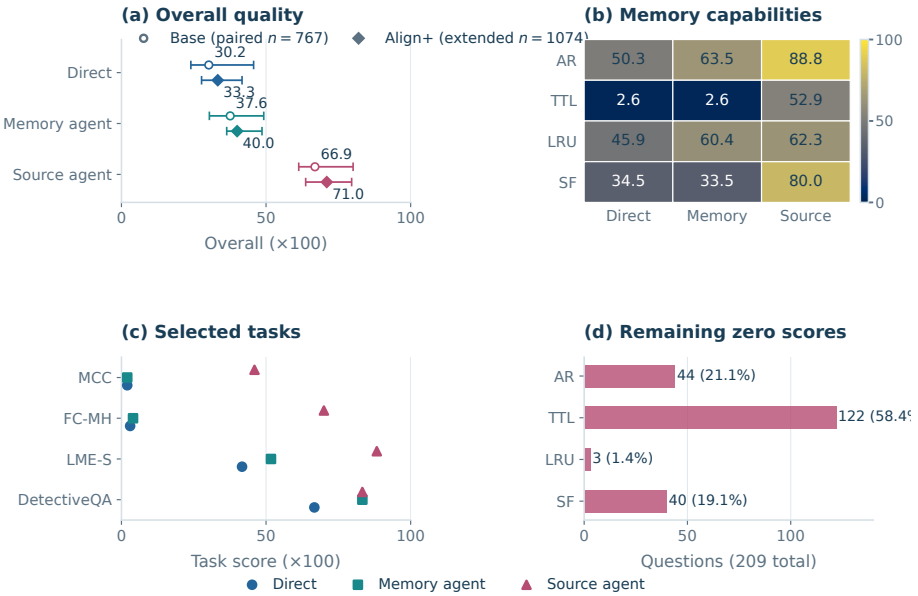

Figure 2: MemoryAgentBench results for fixed retrieval, memory-tool reading, and source-grounded reading. Source reading is strongest on tasks that combine mentions or updates.

4.1 Evaluation setup

We use four public benchmarks. The primary MemoryAgentBench (MAB) evaluation uses a paired 767-question sample over 10 task groups and 59 documents, extended on Align+ to 1,074 questions over 14 task groups (full set: 3,671 questions over 29 task groups); all MAB scores use the official scorer[1] with Kimi-K3 as answerer. For LongMemEval, the complete conversation set is written first, and the three answer methods use paired $n{=}25/25/27$ samples. MAB reports percentages (100 all correct, 0 none); AR is *Accurate Retrieval*, TTL *Test-Time Learning*, LRU *Long-Range Understanding*, SF *Selective Forgetting*; MCC is *Multi-Class Classification*, Recom is movie recommendation scored by Recall@5, and FC-SH/FC-MH are single-hop and multi-hop FactConsolidation; Recall, J, exact-match, and Overall are higher-is-better, while calls, tokens, and bytes are costs. Sampling and judging details are in Appendix A.3, and Appendix Table 4 maps each benchmark to the storage or reading question.

4.2 How to store: updates and entity alignment

LongMemEval asks whether an earlier statement and a later correction can both be found after the complete conversation history has been written. In paired runs, direct retrieval changes from 0.68 to 0.72, memory-tool reading changes from 0.68 to 0.72, and source-grounded reading changes from 0.889 to 0.926 ($n{=}25/25/27$; these tracks are small and directional). Together, these paired runs indicate that old and new observations remain jointly accessible to the read path. The separate 500-question LongMemEval-S run reaches 84.60% with Qwen3.7-plus as both answerer and judge; this protocol differs from the other answerer/judge pairings in the table.

The alignment comparison tests when identity is finalized. Immediate alignment decides window by window; delayed alignment waits for the surrounding evidence. In the paired ingest records, calls fall from 609 to 329 per MAB document and from 63.8 to 22.0 per LongMemEval document. Same-name duplicate families fall from 340 to 124 on MAB and from 482 to 99 on LongMemEval. MCC and FC-MH each increase by 3 points in the paired sample (95% interval -7 to $+13$, including zero). These measurements cover the revised write path, which combines delayed alignment with batched or concurrent processing; its practical effect is lower construction cost with more opportunity to revise identity links. The three-arm breakdown is reported in Appendix Table 5, and Appendix Figure 6 separates its cost, duplicate-family, and quality changes.

MEME separates historical retention from deletion: on 100 episodes (694 after-task checks), deletion accuracy is 0%, cascade 46.34%, and absence 59.23%. The zero is a property of the evaluated write/read path: a MEME deletion event asks the system to stop returning one earlier fact while

[1]The scorer used for these runs is fixed at commit `455306d`.

keeping the surrounding history, and the tested path records that statement as another observation, with no tombstone or query-time suppression rule that propagates from an entity or relation to the affected spans—the evidence is preserved but still retrieved. Supporting deletion requires recording the deletion target, propagating it to derived pointers, and distinguishing current from historical queries. The result locates the missing operation precisely; it is not evidence that source preservation alone implements forgetting. Appendix Figure 8 gives the corresponding update, neighbor-hop, and deletion breakdown.

4.3 How to read: progressive evidence retrieval

Figure 2 gives the panel-level result. Panel (a) rises from about 33 to 40 and then to 71 on Align+; panel (b) shows the source path reaching 88.8 AR, 52.9 TTL, 62.3 LRU, and 80.0 SF while the other paths remain at 2.6 TTL. Panel (c) shows the same advantage on MCC (46.0) and FC-MH (70.0), while panel (d) locates 122 of 209 zero-score questions in TTL. The pattern is consistent with source access and continued evidence gathering for updates and fact consolidation.

The three paths (Section 3.4) differ in what reaches the answerer and who may request more. **Fixed retrieval** materializes one ranked list of memory records and chunks that cannot be changed, reopened, or checked. **Memory-tool reading** may issue several read-only calls, so the second call can use an entity or relation returned by the first, but its context is bounded to returned records with no source-file gate. **Source-grounded reading** selects document groups with the same index, then opens the original files and sends only gate-accepted spans to the answerer. Thus the second path is adaptive navigation over stored records, while the third is adaptive navigation followed by a source-text read.

The table gives the same pattern row by row: source reading is strongest on MCC, FC-MH, MH-QA, and LongMemEval-S, while tool reading mainly helps tasks such as DetectiveQA. These are the cases in which a relevant-looking first chunk is insufficient: the agent must combine mentions, follow an update, or verify an exact qualifier. The fixed path cannot request the missing mention; the tool path can request another bounded record; source reading can inspect the selected file and verify the wording. Appendix Figure 5 and Appendix Table 3 provide the 614-question heatmap and the path-level protocol comparison.

The full-context reference reaches 78.7 on 614 questions whose document groups fit its budget. Source-grounded reading reaches 70.8 on those same questions and 71.0 on the expanded 1,074-question sample; it also remains applicable to the 460 questions that exceed the budget. Full context is strong when it fits, whereas addressed reading continues after the corpus outgrows one window.

LoCoMo tests cross-Episode evidence under the Mem0-compatible binary J (the percentage of answers an LLM judge marks correct). Deep-Dream reaches 95.19% overall (97.38 single-hop, 93.62 multi-hop, 80.21 open-domain, 95.33 temporal), and an independent Qwen3.7-plus cross-judge gives 94.22% on the same answers. We report this as a single-run diagnostic because 1,529 of 1,540 rows predate runtime-code fingerprinting; the published Mem0 and Zep rows retain their own answerers, judges, and test procedures. The multi-hop and temporal results fit the design: relation paths connect Episodes while source observations retain the supporting statements.

A single recorded trace makes the storage–reading boundary observable. The agent first follows concept and relation pointers across two documents, then opens source spans containing \$120, \$225, the quantity 20, and the \$7.50 unit price and computes $120 + 225 + (20 \times 7.50) = 495$. The seven successful calls show how the graph locates candidates while source spans supply the arithmetic; the malformed eighth response is retained as an audit detail. This case study illustrates the control flow and does not replace aggregate benchmark results.

4.4 Comparison with existing memory systems

Table 2 places Deep-Dream beside published results for MAB, LoCoMo, and LongMemEval-S; the model columns make each pairing explicit. In the MAB rows, Kimi-K3 is held fixed: 33.3 with one fixed retrieval result, 40.0 with iterative memory-tool reading, 71.0 with source reading, and the 78.7 full-context row is the same-model upper reference on the 614 questions whose documents fit the window. The published rows retain their reported answerers and judges, so only the same-model MAB pair is a controlled comparison. Category-level reproductions are in Appendix Tables 8–10 and 12. Green rows mark recent strong-model reports: † uses GPT-5-mini/GPT-5, ‡ uses GPT-4.1-mini or GPT-5-mini with GPT-4o-mini judging, and $*$ uses Gemini-3 Pro with GPT-OSS-120B memory and judging (Ji et al., 2026; Wang et al., 2026; Latimer et al., 2025). On LongMemEval-S, Deep-Dream's 84.6 (Qwen3.7-plus answerer and judge) is below MemMachine (93.0), Hindsight (91.4), and Supermemory (85.2): the gaps concentrate in session preference, temporal reasoning, and multi-session questions, while Deep-Dream is strong on assistant facts, user facts, and knowledge

Task / domain	Direct retrieval		Tool-assisted reading		Source reading		Source−Direct (Align+)
	Base	Align+	Base	Align+	Base	Align+	Source−Direct (Align+)
AR: Accurate Retrieval							
SH-QA	54.0	48.0	84.0	72.0	**94.0**	92.0	44.0
MH-QA‡	–	22.0	–	42.0	–	**77.0**	55.0
LongMemEval-S*	26.7	41.7	43.3	51.7	**90.0**	88.3	46.6
EventQA	90.3	89.7	90.7	88.3	**98.0**	**98.0**	8.3
TTL: Temporal and Test-time Learning							
MCC, icl_clinic150	4.0	2.0	3.0	2.0	43.0	**46.0**	44.0
Recom‡	–	3.2	–	3.2	–	**59.8**	56.6
LRU: Long-range Understanding							
∞Bench-Sum†	0.0	0.0	28.4	**29.8**	21.2	0.0	0.0
DetectiveQA	50.0	66.7	50.0	**83.3**	**83.3**	**83.3**	16.6
SF: Selective Forgetting							
FC-SH	68.0	66.0	67.0	63.0	**90.0**	**90.0**	24.0
FC-MH	2.0	3.0	4.0	4.0	67.0	**70.0**	67.0
Overall ($n = 767$)	30.2	32.4	37.6	40.7	**66.9**	65.1	32.7
Overall (n=1074, extended)	–	33.3	–	40.0	–	**71.0**	37.7

Table 1: MemoryAgentBench scores (percent). The last column is Source−Direct within Align+; it is reported when both Align+ paths are available. Bold and underlining mark the best and second-best entry within each row.

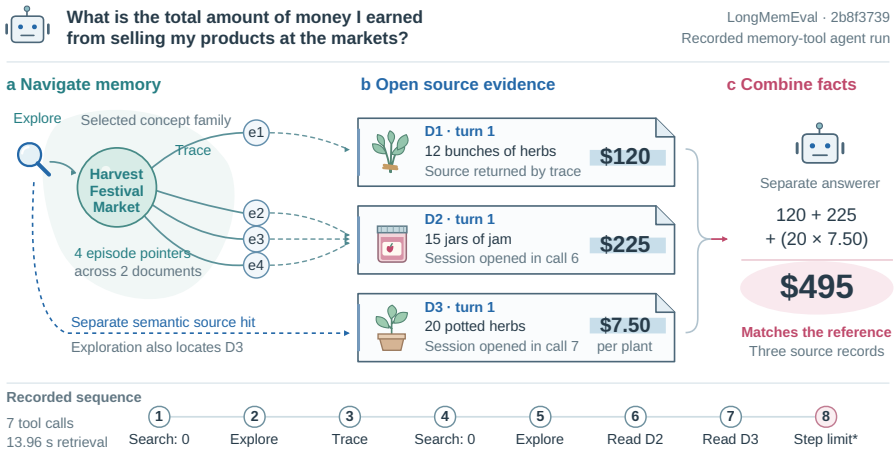

Figure 3: Recorded LongMemEval trace (question `2b8f3739`): concept tracing exposes four episode pointers across two documents, and semantic retrieval locates the third sales record. Seven successful calls read 120, 225, and 7.50; a malformed eighth response reaches the step limit, after which a separate answerer computes 495. The trace shows graph navigation locating candidates and source spans supplying the arithmetic.

updates—the next reading and update targets, with the source-reading strength on MAB and the multi-hop and temporal strength on LoCoMo preserved.

For Infini Memory-A, the MAB row uses the four category averages reported in its paper: 81.2 for AR, 51.0 for TTL, 68.6 for LRU, and 58.0 for SF, with 64.7 overall; its finer subtask scores are not substituted for the category values.

Existing systems use different memory objects: extracted atomic memories (Mem0), episodic graphs with validity windows (Zep/Graphiti), passage graphs (HippoRAG-2), linked notes (A-MEM), or typed modules (MIRIX). Deep-Dream keeps source spans authoritative and uses concepts and relations as revisable addresses, so the agent can verify multi-mention answers in the original files.

(a) MemoryAgentBench (official scorer)

Memory system	Answerer	AR	TTL	LRU	SF	Overall
Full context	GPT-4o	58.1	50.0	54.9	32.5	48.8
Full context	GPT-4o-mini	49.2	48.6	46.2	25.0	42.2
Full context	GPT-4.1-mini	71.8	46.2	49.1	20.5	46.9
Full context	Gemini-2.0-Flash	65.1	46.4	41.6	16.5	42.4
Full context	Claude-3.7-Sonnet	59.7	53.9	62.2	22.5	49.6
Full context (n=614)	Kimi-K3	**96.8**	**75.0**	65.2	78.0	**78.7**
BM25 RAG	GPT-4o-mini	60.5	44.5	35.6	25.5	41.5
HippoRAG-2	GPT-4o-mini	65.1	35.8	36.2	29.5	41.6
MIRIX	GPT-4.1-mini	63.0	35.7	40.5	11.5	37.7
MemGPT	GPT-4o-mini	34.3	40.8	22.4	15.5	28.3
Zep	GPT-4o-mini	37.5	37.5	16.2	5.0	24.0
Mem0	GPT-4o-mini	32.6	21.2	20.7	10.0	21.1
Infini Memory-A	GPT-5-mini†	81.2	51.0	**68.6**	58.0	64.7
Deep-Dream / Direct retrieval	Kimi-K3	50.3	2.6	45.9	34.5	33.3
Deep-Dream / Memory-tool agent	Kimi-K3	63.5	2.6	60.4	33.5	40.0
Deep-Dream / Source reading	**Kimi-K3**	88.8	52.9	62.3	**80.0**	71.0

(b) LoCoMo J score (%; Mem0 scoring)

Memory system	Answerer / judge	Single-hop	Multi-hop	Open-domain	Temporal	Weighted
Mem0	GPT-4o-mini	67.13	51.15	72.93	55.51	≈62.14
Mem0^{g}	GPT-4o-mini	65.71	47.19	75.71	58.13	≈61.36
Zep	GPT-4o-mini	61.70	41.35	76.60	49.31	≈56.32
LangMem	GPT-4o-mini	62.23	47.92	71.12	23.43	≈52.08
OpenAI	GPT-4o-mini	63.79	42.92	62.29	21.71	≈51.10
A-Mem*	GPT-4o-mini	39.79	18.85	54.05	49.91	≈38.95
MemMachine	GPT-4.1-mini / GPT-4o-mini‡	95.12	88.30	71.88	91.59	91.69
Hindsight	Gemini-3 Pro / GPT-OSS-120B*	86.17	70.83	**95.12**	83.80	89.61
Deep-Dream (diagnostic)	Kimi-K3	97.38	93.62	80.21	95.33	95.19
Deep-Dream	Qwen3.7-plus / Qwen3.7-plus	96.91	92.91	77.08	93.46	94.22

(c) LongMemEval-S (%)

Memory system	Answerer / judge	Overall	Sess. pref.	Sess. asst.	Temporal	Multi-sess.	Knowl. upd.	Sess. user
Full context	GPT-4o-mini	55.4	30.0	81.8	36.5	40.6	76.9	81.4
Full context	GPT-4o	60.2	20.0	94.6	45.1	44.3	78.2	81.4
Zep	GPT-4o-mini	63.8	53.3	75.0	54.1	47.4	74.4	92.9
Zep	GPT-4o	71.2	56.7	80.4	62.4	57.9	83.3	92.9
MemMachine	GPT-5-mini‡	**93.0**	**93.3**	98.2	**91.7**	**87.2**	**94.9**	**100.0**
Hindsight	Gemini-3 Pro*	91.4	80.0	96.4	91.0	**87.2**	**94.9**	97.1
Supermemory	Gemini-3 Pro*	85.2	70.0	98.2	82.0	76.7	89.7	98.6
Deep-Dream	**Qwen3.7-plus / Qwen3.7-plus**	84.6	53.3	**100.0**	84.2	74.4	91.0	98.6

Table 2: Results for Deep-Dream and published memory systems on MemoryAgentBench, LoCoMo, and LongMemEval-S. Each row retains its answerer, judge, subset, and evaluation setting. The Deep-Dream LoCoMo row is a single-run diagnostic; its legacy-fingerprint details are in the appendix. Bold marks each column's maximum among the reported comparable rows; the single-run diagnostic row is intentionally unbolded; the Infini Memory-A row uses the category averages reported in its paper.

4.5 Retrieval and evidence diagnostics

We replayed the same 210 questions with four indexes: keyword (BM25) over raw episodes, plus semantic links, plus graph neighbors, plus relation links. Recall-any@10 asks whether one relevant item appears in the top ten; Recall-all@10 asks whether all do. Recall-any@10 is 72.38, 72.38, 71.43, and 72.38%; every paired difference lies within the ±2.4-point 95% bootstrap equivalence bound computed from the same replay, and no other metric moves consistently. The replay therefore gives no evidence that graph links alone lift offline recall; it speaks only to offline retrieval, while online the graph's navigation role—selecting document groups, versions, and relation links before source reading—is exercised, not ablated.

Appendix Figure 7 reports the corresponding call and token costs. Source-grounded reading uses more calls because it verifies and expands evidence, but it also produces the strongest answers on questions that require several mentions or updates.

5 Analysis and Limitations

The main claim is about calibration between storage and reading. A source span is a preserved observation; an identity family, relation, or version is an inferred access structure; a retrieved passage is what the agent elected to inspect. Keeping these categories separate makes revision possible but

does not ensure that an identity is correct or that the inspected passage entails the answer. The source gate validates provenance, not identity or entailment.

The storage choice matters before any query is known. A summary-only store compresses the record in advance, so a later question may need a qualifier, an earlier state, or the original wording that the summary omitted. Deep-Dream keeps that source observation and uses the graph only to make it addressable. A current value therefore remains linked to the observations that produced it, rather than replacing them.

Delayed alignment makes this representation revisable. An early merge can place two people with the same name behind one address and contaminate both histories; a temporary split only adds candidates and can later be corrected by redirecting links. This is why the storage result should be read as a cost and revision result, not as proof that the current identity labels are already correct. Version links preserve how a current value was reached: processing time orders stored states, while dates mentioned in the source remain part of the observation. A query can therefore ask for the latest value or an earlier value without rewriting the record.

Align+ reduces ingest calls by 46–65% while combining delayed alignment, batching, and parallel processing. Same-name families fall from 340 to 124 on MemoryAgentBench and 482 to 99 on LongMemEval. Duplicate-name proxies are not identity labels, and the paired quality interval is -7 to $+13$ points. MCC and FC-MH each change by 3 points (95% interval -7 to $+13$), supporting a construction trade-off rather than a measured identity gain. The entropy argument is conditional on a fixed target and retained predecessor state.

The MAB ladder compares fixed retrieval, memory tools, and source reading at 33.3%, 40.0%, and 71.0% with Kimi-K3 on 1,074 questions. The complete paths differ in tools, source access, iteration, evidence handling, and budgets, so they do not isolate an identity or graph effect. Source reading reaches 70.8% against full context's 78.7% on 614 matched questions; 460 questions exceed the reference budget. The offline graph ablation is null. A smaller context can still cost more if the agent repeatedly searches, and provenance checks do not solve stopping or answer support.

MEME records 0% deletion accuracy on 100 deletion checks within 694 after-task checks, with cascade 46.34% and absence 59.23%. The path lacks target tombstones and propagated suppression. LongMemEval-S reaches 84.6% with Qwen3.7-plus as both answerer and judge under a distinct protocol. LoCoMo's 95.19% and 94.22% are legacy single-run diagnostics because 1,529 of 1,540 rows predate runtime-code fingerprinting. These results call for direct identity labels, entailment checks, and resource-matched reading experiments.

What remains unmeasured. A provenance-checked span can still be irrelevant, and a correctly linked identity can still require another document. The current runs do not provide labeled identity decisions, entailment annotations, or resource-matched stopping curves. Nor does the deletion result test a complete suppression design. These gaps define direct follow-up evaluations: label candidate merges, hold calls and tokens fixed across read policies, and score current and historical queries separately.

6 Conclusion

Long-term memory has two linked problems: preserving a growing record as entities and facts change, and finding only the evidence needed for a question within a finite context window. Deep-Dream addresses storage by keeping document spans factual while concepts, relations, observations, versions, and candidate identities provide revisable access structures. It addresses reading by letting the agent follow those structures to earlier or later passages and stop when the evidence is sufficient.

Each component has a clear role: the graph narrows the search space, the source span preserves the wording and values, and the agent chooses whether another document, relation, or version is needed. The MAB results therefore compare complete reading paths. The write results show that delayed alignment with batched processing lowers ingest cost and duplicate-name families, while the full-context reference remains strongest when its window fits. The MEME result shows why retaining history does not by itself specify when an old fact should be suppressed.

The recorded trace makes this separation concrete: concepts and relations locate candidates, while source spans supply the values. Our contribution is a memory design in which storage and reading can be improved and audited separately, with explicit interfaces for future identity decisions, current-versus-historical queries, entailment checks, and deletion rules.

Ethics Statement

The study uses publicly available benchmark data and recorded software runs. It does not collect personal data, involve human participants, or make decisions about individuals. The reported examples are used only to evaluate memory retrieval and evidence tracking. We follow the licenses and attribution requirements of the benchmark and software sources.

Reproducibility Statement

The released supplementary package contains the run manifests, configurations, prompts, per-item outputs, recomputation scripts, and integrity hashes needed to reproduce the reported tables. Appendix A.1 describes the artifact ledger, while the remaining appendix sections give the sampling frame, model and judge roles, alignment study, and audit details. Results are regenerated from the frozen artifacts rather than entered manually.

AI Use Statement

In this work, we used generative AI tools (including OpenAI Codex) to draft, restructure, and edit the manuscript; to prepare a Chinese reading version; to search for related literature and bibliographic information; to assist with analysis and plotting code; and to formulate the information-theoretic interpretation of the architecture, including mathematical claims and their supporting arguments (progressive entity alignment, search–evidence separation, and query-dependent reading budgets), as well as the memory-hierarchy and addressing framing that interprets context as a working set and the memory store as an addressed structure. Generative image tools were used to create and revise explanatory framework figures. We did not use generative AI tools to produce reported results: no synthetic evaluation dataset was created in this workflow, and all reported benchmark scores are drawn from recorded experimental runs rather than generated estimates.

We have reviewed all AI-assisted work. Reported numbers are regenerated from frozen run artifacts by the released recomputation scripts rather than transcribed from AI output; AI-suggested references were checked against the original publications; and the mathematical statements are presented with their assumptions and tied to the measured values rather than taken as unconditional theorems. AI assistance does not constitute independent proof verification or empirical validation.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pp. 493–507, 1952.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Shaohan Wang, Pengyu Wang, Xiaorui Wang, and Zhendong Mao. A-RAG: Scaling agentic retrieval-augmented generation via hierarchical retrieval interfaces. *arXiv preprint arXiv:2602.03442*, 2026.

Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*, 2025.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pp. 156259–156291, 2026.

Suozhao Ji, Baodong Wu, Zehao Wang, Lei Xia, Qingping Li, Ruisong Wang, Wenbo Ding, Zhenhua Zhu, Boxun Li, Guohao Dai, et al. Infini memory: Maintainable topic documents for long-term llm agent memory. *arXiv preprint arXiv:2606.10677*, 2026.

Seokwon Jung, Alexander Rubinstein, Arnas Uselis, Sangdoo Yun, and Seong Joon Oh. Meme: Multi-entity & evolving memory evaluation. *arXiv preprint arXiv:2605.12477*, 2026.

Chris Latimer, Nicoló Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight is 20/20: Building agent memory that retains, recalls, and reflects. *arXiv preprint arXiv:2512.12818*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. LoCoMo-Plus: Beyond-factual cognitive memory evaluation framework for LLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25085–25100, 2026.

Hai Lin, Hoilam Pao, Shaoxiong Zhan, and Hai-Tao Zheng. Model-native computing architecture: Envisioning future system architecture through the lens of computer architecture. *arXiv preprint arXiv:2606.00288*, 2026.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 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)*, pp. 13851–13870, 2024.

Vasilije Markovic, Lazar Obradovic, Laszlo Hajdu, and Jovan Pavlovic. Optimizing the interface between knowledge graphs and LLMs for complex reasoning. *arXiv preprint arXiv:2505.24478*, 2025.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Claude E Shannon. Coding theorems for a discrete source with a fidelity criterion. In *Claude Elwood Shannon: Collected Papers*. Wiley-IEEE Press, 1993.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Shu Wang, Edwin Yu, Oscar Love, Tom Zhang, Tom Wong, Steve Scargall, and Charles Fan. Memmachine: A ground-truth-preserving memory system for personalized ai agents. *arXiv preprint arXiv:2604.04853*, 2026.

Yu Wang and Xi Chen. MIRIX: Multi-agent memory system for LLM-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. BigCodeBench: Benchmarking code generation with diverse function calls and complex instructions. In *International Conference on Learning Representations*, volume 2025, pp. 66602–66656, 2025.

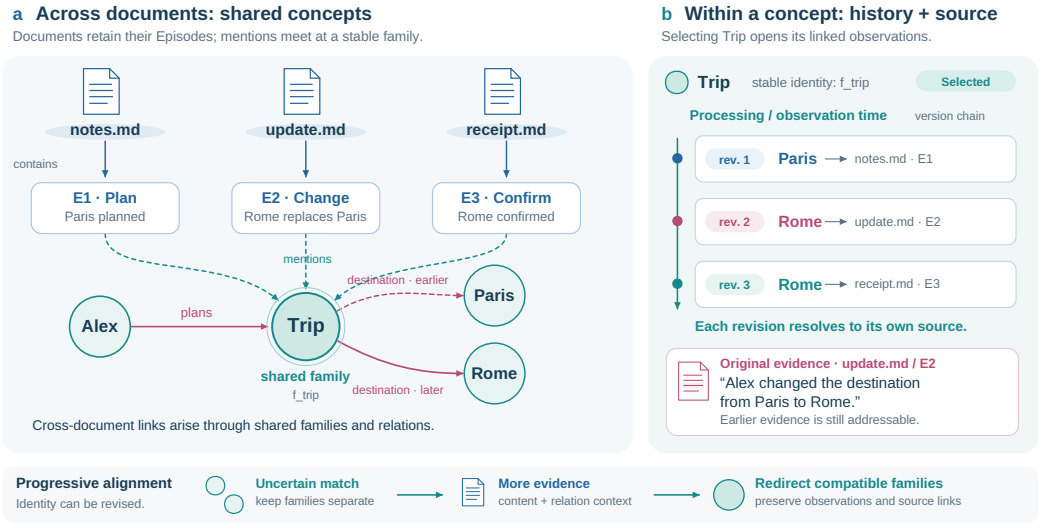

Figure 4: Deep-Dream's document-first temporal concept graph. Multiple documents keep their Episodes and connect through shared concept families and relation edges. Selecting a family exposes its observation-time version chain and the exact source span for each revision; uncertain matches can remain separate and converge when later content and relation context arrive.

A Reproducibility and Audit Details

A.1 Artifact Ledger

The machine-readable ledger is generated by the released ledger and recomputation scripts. It records the source summary, dataset, and per-item JSONL SHA-256 values for every run the paper cites, and is regenerated and compared byte-for-byte with the script output at the submission freeze. The current ledger covers the MemoryAgentBench tracks (two engine generations $\times$ three answer paths, official scorer at commit 455306d, identical `entity2id` fingerprint), LongMemEval complete-write pairs and the 500-question run, LoCoMo-1986/1540 and the K3-as-answerer diagnostic, LoCoMo-Plus-401 (Li et al., 2026), the X7 channel diagnostic, the source-expansion diagnostic, the repaired budget-frontier replay, MEME, BigCodeBench (Zhuo et al., 2025), and ALFWorld (Shridhar et al., 2020). Per-run manifests record dataset and prompt provenance, model identifiers, tool observations, evidence IDs, latency, exit status, and code/runtime hashes; failed integrity checks and negative results are retained rather than dropped. The ledger-generation and recompute scripts ship with the released supplementary material, which pins the SHA-256 of every input too large to redistribute (raw agent trajectories, library snapshots, and the LongMemEval-S dataset file); on the full artifact tree, both scripts regenerate the ledger and the frozen recomputed values byte-for-byte.

The released supplementary material contains the machine-readable manifests (dataset, prompt, model, and runtime provenance), per-item scores and answers, and runtime hashes for every cited run, together with a self-verification script that checks each shipped file against the ledger. The paper reports dataset, model, judge, subset, and integrity status alongside each result so that the comparisons remain auditable without exposing machine-specific storage locations.

Figure 5 compares all four read paths on the same 614 questions. Overall rises from 38.6 for fixed retrieval to 44.6 for tool reading and 70.8 for source reading, while the full-context reference reaches 78.7. The heatmap explains the gap: source reading reaches 95.0 AR and 80.0 SF, close to or above full context, but remains at 46.0 TTL versus 75.0 for full context. The source path therefore recovers most of the retrieval and selective-forgetting signal without loading the entire collection, while update-heavy TTL questions still need stronger temporal handling. Figure 6 separates the bundled Align+ change into its observed effects. Panel (a) shows MAB ingest calls falling from 609 to 329 and LongMemEval calls from 63.8 to 22.0; panel (b) shows duplicate families falling from 340 to 124 and from 482 to 99. Panel (c) records a 3-point increase on both MCC and FC-MH, with the paired 95% interval spanning -7 to $+13$ points. The figure supports a construction-cost and consolidation benefit with compatible sampled quality; because the implementation changes run together, it is not a single-component ablation. The compressed trace appears in Figure 3 in the main text. The released artifact retains the complete tool-call JSON, timestamps, and malformed retry for audit.

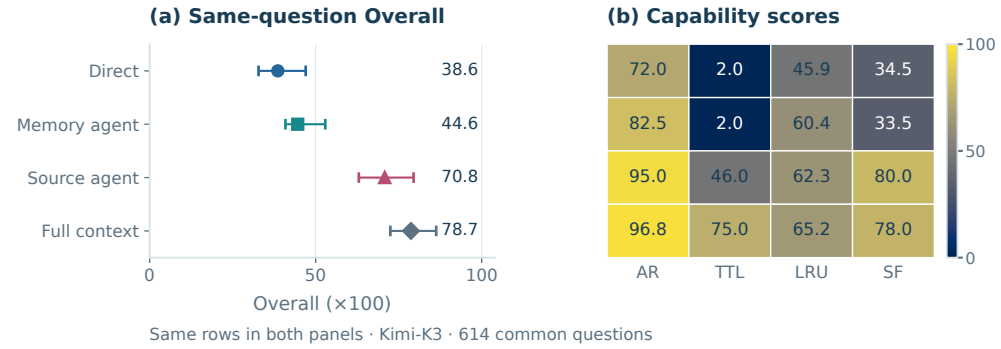

Figure 5: Same-question alignment on the n=614 questions covered by the same-model full-context reference (Kimi-K3 answerer; official macro scorer). (a) Overall scores with marginal 95% bootstrap intervals; the source-reading method scores below the reference by 7.9 points [95% CI −11.9, −3.9]. (b) Domain profiles: the reference's margin concentrates in TTL, while the source-reading method matches or exceeds the reference on AR, LRU, and SF.

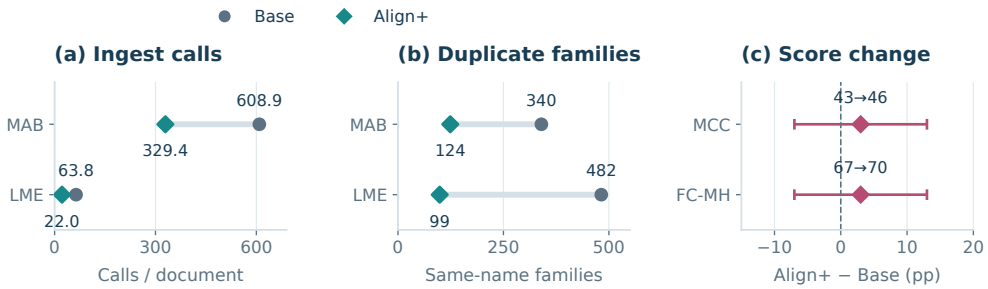

Figure 6: Paired construction diagnostics for the bundled Base→Align+ revision. (a) Ingest calls and same-name duplicate families decrease on the documented MAB and LongMemEval intersections. (b) The paired MCC and FC-MH slices each increase by 3 points; intervals are retained because this is a bundled comparison, not a component ablation.

A.2 Version provenance and alignment A/B study

The accompanying evidence package adds a complete three-arm alignment study on LongMemEval-S: three document groups (42+49+46 documents, 137 documents per arm), write all documents first, and the same Kimi-K3 endpoint. Arm A is the baseline without batch candidate review; B1 enables batch candidate review; B2 additionally enables the revised convergence implementation (parallel alignment, deferred non-destructive merges, and a final consistency check). The saved B1 and B2 service configurations are byte-identical; their difference is in the runtime code and the B2 switch, so the arm assignment is read from the manifest and code snapshot rather than inferred from filenames. The study is powered for efficiency and consolidation, not quality: each document group contributes only one question per evaluation track.

The separate LME Base/Align+ engine diff (initial/revised implementation) is a post-hoc reconstruction from commits $73d9c5e$ to $8fa9651$. The latter also contains four throughput fixes, and therefore cannot by itself establish which fixes were active at each point in the Align+ run. We treat the manifest-embedded pipeline configuration, runtime-policy hash, and skill hash as authoritative; the reconstructed diff is provenance for the uncommitted code, not a claim that every later fix was active from the first revised run.

A.3 MemoryAgentBench sampling frame and contamination note

The sampled subset was drawn with a fixed task-group list recorded in the released ledger: a 767-question paired sample over 10 task groups and 59 documents, extended in place on the Align+ engine with four task groups (RULER-421K multi-hop QA, 100 questions; the full Redial recommendation corpus, 200 questions; one further DetectiveQA task group, 6 questions; one further ∞Bench-Sum task group, 1 question) to 1,074 questions over 14 task groups, against the full 3,671 questions over 29 tasks. All four competencies are covered; 17 tasks remain unsampled after the extension. The paired sample and its `entity2id` fingerprint are shared by all six runs (three paths × two generations);

Property	Direct retrieval	Tool-assisted reading	Source reading
Source access	one ranked payload	episode/session payloads	selected source files
Tools	none	memory tools	scope, grep, script
Answerer	Kimi-K3	Kimi-K3	Kimi-K3
Loop	none	several read-only calls	several read-and-cite calls
Cost / question	27.5k tokens, 1.0 LLM calls	43.2k tokens, 6.5 LLM calls	328.6k tokens, 10.2 LLM calls

Table 3: Differences between the three reading methods. The source path adds a deterministic source check that confirms returned IDs resolve to source offsets; the final answerer remains Kimi-K3; all paths share frozen libraries, question stream, and official scorer.

Line	Capability	Benchmarks and observables	Caliber and boundary
Store	Lossless accumulation and revisable identity	LongMemEval complete-write paired runs (n=25/25/27) and the 500-question LME-S run test whether an earlier statement and a later correction stay jointly reachable; Base/Align+ paired runs measure same-name duplicate families (340→124 on MAB, 482→99 on LME) and ingest calls (609→329 per MAB document; 63.8→22.0 per LME document), with MCC/FC-MH deltas whose intervals include zero.	MEME: 0% deletion real accuracy on 100 episodes (694 after-task checks); cascade/absence real accuracy 46.34/59.23; no query-time deletion suppression yet.
Read	Agent-owned addressing	MemoryAgentBench compares one fixed retrieval result, repeated read-only tool calls, and citation-checked source reading on the same 767-of-3,671-question paired sample (8 of 29 task groups), with the Align+ extension to 1,074 questions; LoCoMo-1540 binary J separates single-hop, multi-hop, temporal, and open-domain questions.	Published Mem0/Zep rows use different answerers, judges, and rubric revisions—background only; the channel replay is an offline equivalence bound ($\pm$2.4 points, 95% bootstrap), not an online ablation.

Table 4: Two research questions and their measurements. Each row groups benchmarks by the line they probe, with calibers and disclosed boundaries adjacent.

the 307 extension questions exist on Align+ only, and the two large extension document groups (1.9M and 5.7M characters) exceed the full-context reference's strict budget of 1,042,576 characters, so reference coverage is 614 of the 1,074 questions. The reference uses the neutral prompt profile; the normalized-v1 profile of the same run scores 46.0 on its own 607-question coverage, a prompt-sensitivity artifact retained in the ledger and not used as a reference. Document-group lengths are heterogeneous and disclosed wherever cross-system cells lead: the two FactConsolidation document groups are the 6K variants (26,157 characters each); the 32K/64K/262K FC variants (136k–1.1M characters) and the other unsampled document groups were dropped for budget, all listed with context sizes in the released ledger's subset block; EventQA is sampled at the 64K/128K/full-book document groups rather than only the published full-book average; and the LRU domain carries n=7 paired and 14 extended questions (∞Bench-Sum 1 then 2, DetectiveQA 6 then 12), so LRU cells are directional. The MAB source corpora are public; whether the answerer's pretraining saw them cannot be excluded or measured post hoc, and every cross-system table row states its own answerer, judge, and subset.

A.4 Adjusted Overall derivation

The paired sample omitted two tasks that score low for everyone, AR's MH-QA and TTL's Recommendation, and an earlier draft handled this by imputation: entering published non-full-context reference values (Recommendation ≈13, MH-QA ≈60) gave an adjusted Overall of ≈59.0 against the raw 65.1 at the paired caliber, with a 59–61 range under sensitivity. The extension supersedes the imputation: both tasks are now measured directly on Align+ (source-reading MH-QA 77.0, Recommendation 59.8; Table 1), and the extended Overall is 71.0 with both tasks inside the aggregation. The extended sample's composition still differs from the full benchmark's (17 tasks remain unsampled), so we continue to state raw values only and never subtract across calibers; the imputed estimate is retained here as provenance for the released ledger.

A.5 Per-path prompts and configurations

All paths use Kimi-K3 as answerer, with their prompts archived verbatim in the run manifests and code snapshots listed in Appendix A.1; the released supplementary material reproduces them in full. The Base→Align+ engine configurations differ as follows: max output tokens 16,384→32,768; usable context window 32k→262,144; retries 2→5; timeout 300 s→1,200 s; and absolute-path persistence enabled in Align+. Direct retrieval makes one fused retrieval call and one answer call; the

Arm	Alignment revision	Calls/doc	Total tokens/doc	Entity redirects	Duplicate groups
A	Baseline	68.4	145,378	0	8
B1	Window-batch alignment	24.2	108,467	1	12
B2	Batch review + final consistency check	29.4	111,853	245	6

Table 5: Three-arm alignment study on 137 LongMemEval-S documents per arm. B1 provides the large efficiency reduction; B2 spends about 3% more tokens than B1 on final consistency check and is the only arm with substantial non-destructive family redirection, reducing same-name duplicate groups. The quality slice is too small for a quality claim. B1 and B2 use the same saved service configuration; B2 differs through the runtime implementation and alignment switch.

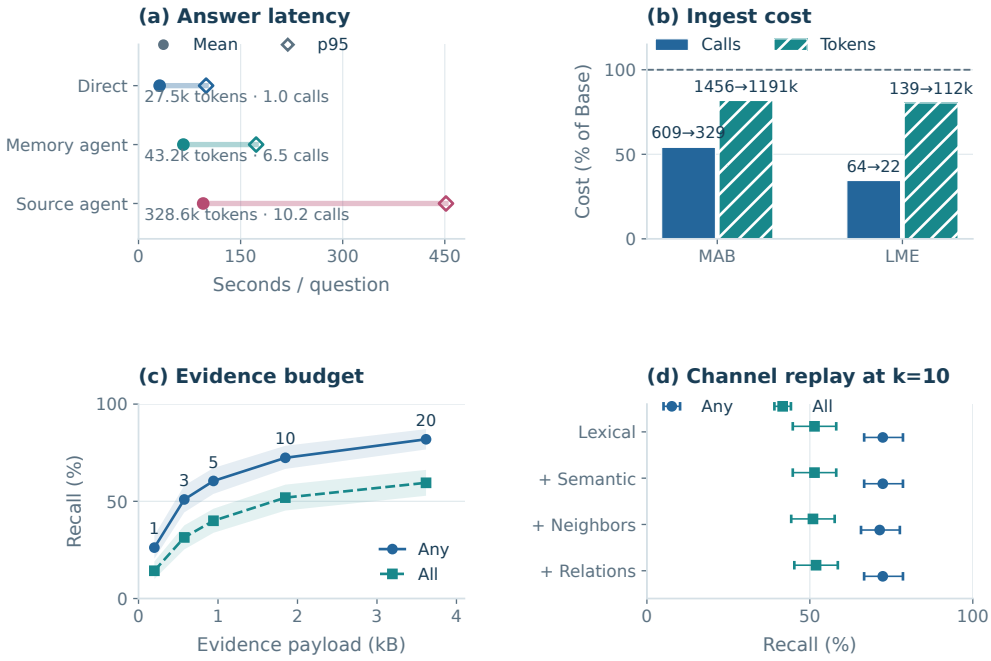

Figure 7: Operating-point diagnostics from frozen artifacts (Align+ engine; panel a over all 1,074 questions). (a) Mean-to-p95 answer latency with tokens and calls annotated; (b) paired Base→Align+ ingest cost; (c) the valid fixed-ranking retrieval-budget frontier; (d) source-only channel replay at k=10.

Memory-tool agent uses a repeated read-only calls to the tool; the Source-reading agent materializes a document group, reads and verifies source files, and submits only spans that pass the citation check.

Figure 7 links answer quality to operating cost. Panel (a) gives mean-to-p95 latency of 31.0–99.3 seconds for fixed retrieval, 65.9–172.8 seconds for tool reading, and 95.0–451.7 seconds for source reading; the source path spends 10.2 calls and 328.6k tokens per question. Panel (b) shows why delayed alignment matters operationally: calls fall 609→329 and 64→22, while tokens fall 1,456k→1,191k and 139k→112k. Panel (c) gives the budget trade-off directly: increasing k from 1 to 20 raises Recall-any from 26.2% to 81.9% and Recall-all from 14.3% to 59.5%, while the mean payload grows from 0.2 to 3.6 kB. Panel (d) is comparatively flat across lexical, semantic, neighbor, and relation channels, showing that the larger gains come from expanding the read path rather than changing one offline channel.

A.6 Additional quantitative diagnostics

The ledger also contains type-level update scores, a controlled source-expansion replay, and the MEME deletion boundary. Figure 8 reports these quantities with their sample-size and development-set limits visible.

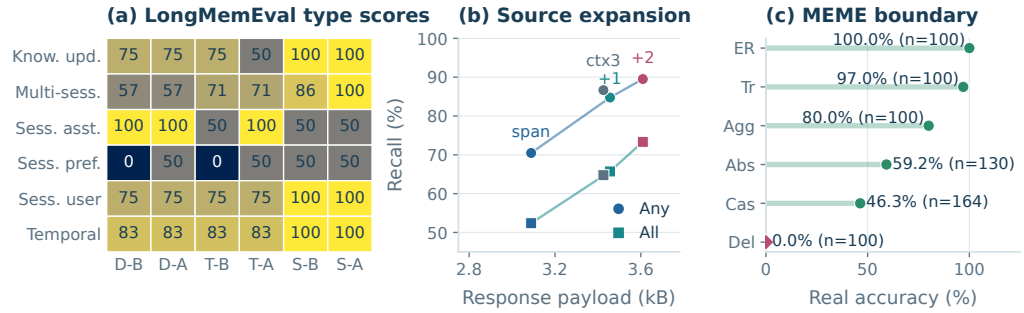

Figure 8: Additional quantitative diagnostics from the frozen ledger. (a) LongMemEval type scores for the paired Base/Align+ tracks; the directional samples have $n{=}25/25/27$ and are shown as coarse percentages. (b) Source-expansion recall versus response payload on 105 prior-error cases; this is a development diagnostic, not a held-out test. (c) MEME real task accuracy with task-specific denominators, exposing the 0% deletion boundary.

Figure 8 gives three checks on the mechanism. Panel (a) shows the clearest LongMemEval change in multi-session access: source reading rises from 85.7% to 100% between Base and Align+, while knowledge-update and temporal scores remain at 100%; this is consistent with better cross-session consolidation rather than a uniform score shift. Panel (b) shows that adding one and two neighbor hops increases Recall-any from 70.5% to 84.8% and 89.5%, and Recall-all from 52.4% to 65.7% and 73.3%, at a payload increase of about 0.4–0.5 kB. Panel (c) separates task semantics: entity resolution is 100%, tracking 97%, aggregation 80%, absence 59.2%, cascade 46.3%, and deletion 0%. Thus the graph supports locating related evidence, while deletion still requires an explicit suppression operation.

A.7 Carrier anchors (BigCodeBench (Zhuo et al., 2025), ALFWorld (Shridhar et al., 2020))

BigCodeBench: 0.4859 completion / 0.2838 hard pass@1 (calibrated split). ALFWorld: 0.9786 in-distribution / 0.9851 out-of-distribution success. Both are memory-insensitive agent-carrier regressions: they anchor that the agent substrate executes tasks correctly, and are not compared horizontally with memory benchmarks.

A.8 MAB ingest manifest gaps

The Base and Align+ MAB ingest runs share a 45-document paired intersection. Manifests missing from one generation are excluded from per-document efficiency ratios, which are therefore computed on the intersection only (609→329 calls/doc, -46%); token ratios use the same intersection. Question-level results are unaffected: all 767 paired questions run against frozen libraries in both generations, and the 307 extension questions ran on the Align+ engine only, so ingest-cost ratios remain defined on the paired intersection.

A.9 Budget frontier

An earlier depth-10-versus-depth-20 replay is retained as an integrity finding: its prefix invariant failed on all 210 questions (`primary_result_invariant_violations=210`), so it is not used for any positive claim. The repaired design runs retrieval once per query at maximum depth and slices a fixed ranking to $k \in \{1, 3, 5, 10, 20\}$; prefix monotonicity then holds by construction (zero violations), and the frontier rises monotonically from $k{=}1$ to $k{=}20$ with per-k percentile-bootstrap CIs recomputed from the frozen replay artifacts.

A.10 X7 channel ablation detail

A.11 Protocol boundaries

The original LoCoMo token-F1 track includes 1,986 questions and five categories. The Mem0-compatible track contains 1,540 questions and four categories. LongMemEval-S contains 500 ques-

Table 6: X7 provenance ablation, offline retrieval replay on the 210-question gate set. Arms 1–4 form a clean channel ablation: only `enabled_channels` differs, so any recall delta is attributable to that channel set. Arms 5 (evidence-bound agent, gate ON) and 6 (unsurfaced evidence, gate relaxed) reuse arm 4's five-channel ranking and are retrieval-identical to arm 4; their distinguishing columns (QA accuracy, unsupported-answer rate, mean evidence tokens) require the LLM agent run and are pending. The evidence-gate plumbing was proved offline: with the gate ON, submitting an unsurfaced turn is rejected; with the gate relaxed the same submission is accepted (`plumbing_ok`=true). In the deployed MAB runs the gate fired as designed: 14/1074 questions in the deployed $v2$ skill-agent run had an evidence submission rejected as unsurfaced (9/767 in $v1$); 11 of the 14 resubmitted tool-surfaced evidence inside the loop and the remaining 3 ended at the step cap and were answered from tool-surfaced context; zero rejections on the LME runs. Bootstrap CIs: 2,000 resamples, 95%.

Arm	Channels (additive)	Recall-any@10 [95% CI]	Recall-all@10 [95% CI]	Mean bytes@10
1	raw-doc + episode-bm25	72.38 [66.67, 78.57]	51.43 [44.76, 58.10]	1833
2	+ semantic-provenance	72.38 [66.67, 78.57]	51.43 [44.76, 58.10]	1841
3	+ graph-neighbor	71.43 [65.70, 77.62]	50.95 [44.29, 57.62]	1845
4	+ relation-evidence (full 5)	72.38 [66.65, 78.57]	51.90 [45.24, 58.58]	1847

Benchmark	Protocol	Score	Answerer / judge
LongMemEval-S (500)	sampled, n=500	84.60	Qwen3.7-plus answerer; Qwen3.7-plus judge
LongMemEval, direct	paired, n=25	0.68 → 0.72	Kimi-K3 answerer
LongMemEval, tool-assisted	paired, n=25	0.68 → 0.72	Kimi-K3 answerer
LongMemEval, source-reading	paired, n=27	0.889 → 0.926	Kimi-K3; self-judged
LoCoMo-1540	Mem0 J scoring	95.19	Kimi-K3; answerer+judge; n=1
LoCoMo-1540 (same answers)	cross-judged	94.22	Qwen3.7-plus; agreement 98.51%

Table 7: Results on other benchmarks under their stated test procedures; each row states its answerer and judge. The LoCoMo run differs from Mem0's published J in two admitted ways—single run versus 10-run mean±std, and the answerer/judge model—and carries a disclosed legacy-fingerprint caveat; we state these conditions so the comparison is reproducible. The LME source-reading agent track's two extra questions follow its session evidence rule; comparisons are made only within the same track.

tions; the complete-write paired tracks use n=25/25/27 as described in Section 4.1. MEME's 100 episodes produce heterogeneous task counts (ER/Agg/Tr/Del/Cas/Abs = 100/100/100/100/164/130 after-task checks in the cited run). Raw and real scores are not interchangeable.

A.12 Integrity failures retained

The depth replay of Appendix A.9 reports `primary_result_invariant_violations=210` and `passed=false`. The MEME deletion result is 0% real accuracy. The X7 channel ablation is a null result, reported as an equivalence bound rather than dropped. The ∞Bench-Sum n=1 flip costs the Align+ Overall 2.7 points directly (the net Base→Align+ gap is 1.8 points) and is disclosed at every comparison. The LME Direct retrieval–Memory-tool agent tie is reported although it is unflattering to the Memory-tool agent path. These records remain visible because an auditable memory paper should expose failed and null mechanisms as well as successful ones.

A.13 External comparison test details

Mem0's original paper reports LoCoMo category-level F1, BLEU-1, and a binary J in which a GPT-4o-mini judge labels each answer CORRECT or WRONG and the score is the percent judged CORRECT (mean±std over 10 independent runs). Zep's original paper reports DMR and LongMemEval with GPT-4o-family answerers. Rather than re-run those systems under our protocol, we reproduce their published tables verbatim below (Tables 8–10) as references reported under different test procedures. The Mem0 deterministic token-overlap metrics omitted from the main tables for space are reproduced in Table 12.

The K3-substitution row. The single Deep-Dream J row placed inside Mem0's J column is produced by running Mem0's exact binary scoring procedure with Kimi-K3 substituted for GPT-4o-mini as *both* the answerer and the judge—the only change is the model. This yields 95.19% under the same-model K3 judge (97.38/93.62/80.21/95.33 S/M/O/T) and 94.22% under an independent Qwen3.7-plus cross-judge on the same K3 answers (96.91/92.91/77.08/93.46; cross-judge agreement 98.51%). Two differences from Mem0's published J are admitted: it is a single run rather than a 10-run mean±std, and the model is Kimi-K3 rather

than GPT-4o-mini. We additionally disclose an integrity caveat: the run manifest carries `diagnostic_override={kind:legacy-missing-runtime-code-fingerprint, formal_claim_allowed:false}`, because 1,529 of the 1,540 answer rows originate from a legacy source run (`legacy-source-run`) that predates runtime-code fingerprinting. The answers themselves are unchanged (`answer_rows_unchanged=true`, answer SHA-256 76af5772) and the per-item JSONL carries `runtime_code_sha256` on all 2,063 rows (3 distinct hashes: 2fa32774 on 2,019, 36b45933 on 33, 42afb1ac on 11), the same dataset SHA-256 (79fa87e9) and library snapshot (cf2e76b8) as the Qwen-answerer run, and the same Mem0 binary-judge prompt (repository `mem0ai/memory-benchmarks`, commit 4b61c5d, mode `unified-no-evidence-exact-direct`). We therefore report 95.19% as a diagnostic cross-judged result, not a fingerprint-complete formal entry.

A stricter controlled comparison—freezing the same answer model, judge model, prompt, temperature, evidence budget, dataset revision, and Graphiti/Zep implementation commit—remains the only path to a shared leaderboard, but it is not a submission hard gate now that the published numbers are cited directly.

A.14 Formal properties of storage and reading

This section makes the storage and reading claims precise. The first result is a structural property of versioned, append-only storage. The second is a conditional statistical statement about delayed alignment. The remaining equations describe the read policy and its answer-risk decomposition.

A.14.1 Write time: state updates and progressive alignment

Let $\mathcal{D}$ be the set of source documents. For $D \in \mathcal{D}$, let $\mathcal{E}(D)$ be its Episodes. The ith observation is

$$o_i = (D_i, e_i, \tau_i, x_i, \ell_i), \qquad e_i \in \mathcal{E}(D_i),$$

with processing time τ_i, extracted content x_i, and immutable source offset ℓ_i. Let $O_t = \{o_1, \dots, o_t\}$ and $M_t = (F_t, O_t, G_t)$ be the state after t observations. The abstract write operator used in the proof is

$$\begin{aligned} F_{t+1} &= \mathsf{Align}(F_t, x_{t+1}), \\ O_{t+1} &= O_t \cup \{o_{t+1}\}, \\ G_{t+1} &= \mathsf{Link}(G_t, F_{t+1}, o_{t+1}), \\ M_{t+1} &= (F_{t+1}, O_{t+1}, G_{t+1}). \end{aligned} \tag{5}$$

The operator Align may create a candidate family or merge compatible families; Link adds or redirects pointers. Neither operator changes the source addressed by ℓ_i. The version chain supplies a projection π_t such that $\pi_t(M_{t+1}) = M_t$; this is the formal statement that a later state retains the earlier state for audit and revision.

Proposition A.1 (Revisability of delayed alignment)**.** *Let $\mathcal{P}_t$ be the partition of O_t induced by the candidate families in F_t. For every coarsening $\mathcal{P}'_t$ of $\mathcal{P}_t$, there is a non-destructive pointer map that realizes $\mathcal{P}'_t$ without changing O_t or any source offset.*

Proof. A coarsening assigns each old family f to a new family $m(f)$. Keep every observation $o_i = (D_i, e_i, \tau_i, x_i, \ell_i)$ unchanged and replace each family pointer by $m(f)$. Replace an edge (f, r, o_i) in G_t by $(m(f), r, o_i)$. The observation set and every ℓ_i are therefore identical before and after the operation, while observations formerly in different blocks can now be reached through the same family. This realizes $\mathcal{P}'_t$ without destructive rewriting. A split is also safe: it copies pointers into separate candidate families while leaving O_t unchanged. If a system instead stores only a deterministic summary $\sigma = g(O_t)$, then $Z \to O_t \to \sigma$ is a Markov chain for the latent entity assignment Z. The data-processing inequality gives $I(Z;\sigma) \le I(Z;O_t)$; the inequality can be strict when g discards information needed to distinguish two identities. □

Proposition A.2 (Convergence with pooled evidence)**.** *Assume that, conditional on entity identity, k feature observations are independent. Let two entities $e \neq e'$ have distinguishable densities p_e and $p_{e'}$ with Chernoff coefficient below one. Define their Chernoff coefficient and information by*

$$\rho_{e,e'} = \inf_{0 \le s \le 1} \int p_e(x)^s p_{e'}(x)^{1-s}\, dx, \qquad C_{\mathrm{Ch}}(e,e') = -\log \rho_{e,e'} > 0.$$

For the equal-prior likelihood-ratio test of e versus e', the Bayes error after k pooled observations satisfies (Chernoff, 1952)

$$\Pr[\text{error}] \le \frac{1}{2}\, \rho_{e,e'}^k = \frac{1}{2} \exp[-kC_{\mathrm{Ch}}(e,e')]\,. \tag{6}$$

Proof. For equal priors, the Bayes error is

$$\Pr[\text{error}] = \frac{1}{2}\int \min\left\{\prod_{j=1}^{k} p_e(x_j), \prod_{j=1}^{k} p_{e'}(x_j)\right\} dx_{1:k}.$$

For $0 \le s \le 1$, $\min\{a,b\} \le a^s b^{1-s}$. Applying this pointwise and using independence gives

$$\Pr[\text{error}] \le \frac{1}{2}\prod_{j=1}^{k}\int p_e(x_j)^s p_{e'}(x_j)^{1-s}\, dx_j = \frac{1}{2}\left(\int p_e(x)^s p_{e'}(x)^{1-s}\, dx\right)^k.$$

Taking the infimum over s yields Equation (6). The bound describes an ideal test; real extraction noise and correlated windows determine how closely an implementation approaches it. □

Thus, delayed alignment has a clear qualitative prediction: additional observations can make a correct merge easier to recognize, while an early wrong merge can contaminate one address with two histories. The measured duplicate-family counts (340→124 on MAB and 482→99 on LongMemEval) and ingest-call reductions (46–65%) are system measurements, not estimates of C_{Ch}.

A.14.2 Read time: hierarchical search and step-by-step reading

Let $\mathcal{S}_t$ be the set of immutable source spans reachable from M_t. Given question q, set $C_0(q) = \varnothing$ and let

$$a_{k+1} = \rho\big(q, M_t, C_k(q), u_k\big) \in \big(\mathcal{S}_t \setminus C_k(q)\big) \cup \{\mathsf{stop}\}.$$

For a span action, set $C_{k+1}(q) = C_k(q) \cup \{a_{k+1}\}$; for stop, keep $C_{k+1} = C_k$ and terminate. Induction on k proves $\varnothing = C_0(q) \subseteq C_1(q) \subseteq \cdots \subseteq \mathcal{S}_t$: the base case is immediate, and the induction step adds either one member of $\mathcal{S}_t$ or nothing. The equation tracks source-opening actions; index-only lookups update the policy state u_k without changing C_k. This is a property of the read protocol; it does not assume that every selected span is relevant.

Let $A(q)$ be the compact concept/relation record used to choose candidates, E_C the exact source spans opened from C, and Y the answer-relevant fact. The chain rule is

$$\begin{aligned} I(Y; A, E_C \mid Q=q) &= H(Y \mid Q=q) - H(Y \mid A, E_C, Q=q) \\ &= \big[H(Y \mid Q=q) - H(Y \mid A, Q=q)\big] \\ &\quad + \big[H(Y \mid A, Q=q) - H(Y \mid A, E_C, Q=q)\big] \\ &= I(Y; A \mid Q=q) + I(Y; E_C \mid A, Q=q). \end{aligned} \tag{7}$$

The first bracket is the information used to locate candidates; the second is the additional information obtained by opening source text. If a summary-only system replaces the source collection S by $\widetilde{S} = f(S)$, then, conditional on Q, the Markov chain $Y \to S \to \widetilde{S}$ gives $I(Y; \widetilde{S} \mid Q) \le I(Y; S \mid Q)$ by data processing. Deep-Dream keeps S available after the address layer has narrowed the search; it does not claim to maximize either mutual information term.

For the answer-risk decomposition, let R_q be the event that a relevant source span is in the final opened set $C_K(q)$, and let $V_q = \{\widehat{Y} = Y\}$ be the event that the answer is correct. The law of total probability gives

$$\begin{aligned} \Pr[V_q \mid q] &= \Pr[R_q \mid q]\Pr[V_q \mid R_q, q] \\ &\quad + \Pr[\neg R_q \mid q]\Pr[V_q \mid \neg R_q, q] \\ &\ge \Pr[R_q \mid q]\Pr[V_q \mid R_q, q]. \end{aligned} \tag{8}$$

The first factor is affected by concept and relation navigation; the second is affected by reading and judging the opened evidence. This inequality is a bookkeeping decomposition, not an empirical accuracy bound.

Proposition A.3 (One-sided effect of an address error)**.** *Assume the citation gate accepts a citation c only when c is returned by the current read path and its offset resolves to the immutable source text. An error in F_t or G_t can change which spans enter $C_K(q)$, but cannot change the text or offset of an accepted citation.*

Table 8: LoCoMo per-category J (binary CORRECT/WRONG accuracy, LLM-judged). Published rows reproduced verbatim from Chhikara et al. (2025) (arXiv:2504.19413, Table 1): Mem0 reports mean±std over 10 runs with GPT-4o-mini as both answerer and judge (std omitted, all $\leq$0.75). **Deep-Dream** applies the binary J rubric pinned at repository commit `4b61c5d`—a 2026 revision of the Mem0 judging prompt whose tolerance rules differ from the 2025 paper's Appendix A—with Kimi-K3 as both answerer and judge (single run); the † row judges the same K3 answers with an independent Qwen3.7-plus cross-judge (cross-judge agreement 98.51%). Same 841/282/96/321 question populations (1,540 total, Mem0-compatible subset). *Caveat:* the K3-answerer run is flagged `formal_claim_allowed=false`—1,529 historical answer rows predate runtime-code fingerprinting (answers unchanged, sha 76af5772; per-item JSONL carries `runtime_code_sha256`); reported as a diagnostic cross-judge result, not a fingerprint-complete formal result. Deterministic F1/BLEU-1 omitted for space (Table 12). Weighted J, recomputed from the published category means under the 841/282/96/321 counts: Mem0 ≈62.14, $Mem0^g$ ≈61.36; the paper's own Table 2 reports Overall $J = 66.88$ (Mem0) and 68.44 ($Mem0^g$). Deep-Dream agg. J: 95.19% / 94.22% (†).

Method	J (S/M/O/T)
LoCoMo	— / — / — / —
ReadAgent	— / — / — / —
MemoryBank	— / — / — / —
MemGPT	— / — / — / —
A-Mem	— / — / — / —
A-Mem*	39.79 / 18.85 / 54.05 / 49.91
LangMem	62.23 / 47.92 / 71.12 / 23.43
Zep	61.70 / 41.35 / 76.60 / 49.31
OpenAI	63.79 / 42.92 / 62.29 / 21.71
Mem0	67.13 / 51.15 / 72.93 / 55.51
$Mem0^g$	65.71 / 47.19 / 75.71 / 58.13
Deep-Dream	97.38 / 93.62 / 80.21 / 95.33
Deep-Dream†	96.91 / 92.91 / 77.08 / 93.46

Table 9: Published Deep Memory Retrieval (DMR) baselines reproduced verbatim from Rasmussen et al. (2025) (arXiv:2501.13956). DMR has 500 multi-session conversations (5 sessions, $\leq$12 messages each) with one single-hop fact-retrieval question per conversation. Deep-Dream does not currently run DMR, so this table is a contextual reference only.

Method	Answerer	Accuracy
Recursive Summarization	GPT-4-turbo	35.3%
Conversation Summaries	GPT-4-turbo	78.6%
MemGPT	GPT-4-turbo	93.4%
Full-conversation	GPT-4-turbo	94.4%
Zep	GPT-4-turbo	94.8%
Conversation Summaries	GPT-4o-mini	88.0%
Full-conversation	GPT-4o-mini	98.0%
Zep	GPT-4o-mini	98.2%

Proof. The gate's acceptance predicate is

$$\mathsf{Accept}(c) = \mathbf{1}\{c \in C_K(q)\}\mathbf{1}\{\mathsf{resolve}(c) = \mathrm{source}(D, \ell)\}.$$

The first factor depends on the derived address layer and may be zero for a relevant span. The second factor reads the immutable pair (D, ℓ) and is unchanged by edits to F_t or G_t. Hence an address error can cause missing evidence, but cannot substitute new text for an accepted source span. The proposition does not guarantee that the accepted span semantically entails the answer. □

The conditional results above explain the design measurements without turning them into unmeasured causal claims: the Chernoff statement is an idealized prediction, duplicate-family counts are a proxy for consolidation, and the benchmark results remain the empirical test of the complete system.

Table 10: LongMemEval main results. Published rows (GPT-4o-mini / GPT-4o, Full-context vs. Zep) reproduced verbatim from Rasmussen et al. (2025) (arXiv:2501.13956). The **Deep-Dream** row uses Qwen3.7-plus as answerer and Qwen3.7-plus as judge—a different test procedure; its accuracy (84.6%) is shown beside the published numbers with that condition stated. Latency and average context for Deep-Dream were not measured under the published test procedure (—). The six-category breakdown is in Table 11.

Answerer	System	Accuracy	Latency (s)	Avg. context
GPT-4o-mini	Full-context	55.4%	31.30	115k
GPT-4o-mini	Zep	63.8%	3.20	1.6k
GPT-4o	Full-context	60.2%	28.90	115k
GPT-4o	Zep	71.2%	2.58	1.6k
Qwen3.7-plus	**Deep-Dream**	**84.6%**	—	—

Table 11: LongMemEval six-category breakdown (binary accuracy, %). Published Full-context and Zep values reproduced verbatim from Rasmussen et al. (2025) (arXiv:2501.13956); the **Deep-Dream (Q.)** row is Qwen3.7-plus answerer+cross-judge on the same six categories. Different answerer/judge, so this is a matched-granularity contextual reference, not a shared leaderboard. Sess. pref. is the hardest category for all systems.

System	Sess. pref.	Sess. asst.	Temporal	Multi-sess.	Knowledge upd.	Sess. user
Full-ctx (4o-mini)	30.0	81.8	36.5	40.6	76.9	81.4
Zep (4o-mini)	53.3	75.0	54.1	47.4	74.4	92.9
Full-ctx (4o)	20.0	94.6	45.1	44.3	78.2	81.4
Zep (4o)	56.7	80.4	62.4	57.9	83.3	92.9
Deep-Dream (Q.)	**53.3**	**100.0**	**84.2**	**74.4**	**91.0**	**98.6**

Table 12: Mem0 LoCoMo deterministic metrics, reproduced verbatim from Chhikara et al. (2025) (Table 1). F1 / BLEU-1 over the four categories (Single-hop / Multi-hop / Open-domain / Temporal). Deep-Dream does not compute token-overlap metrics (its metric is binary LLM-judge accuracy, Table 8), so no Deep-Dream column appears here; this table is included for completeness of the published reproduction only.

Method	F1 (S/M/O/T)	BLEU-1 (S/M/O/T)
LoCoMo	25.02 / 12.04 / 40.36 / 18.41	19.75 / 11.16 / 29.05 / 14.77
ReadAgent	9.15 / 5.31 / 9.67 / 12.60	6.48 / 5.12 / 7.66 / 8.87
MemoryBank	5.00 / 5.56 / 6.61 / 9.68	4.77 / 5.94 / 5.16 / 6.99
MemGPT	26.65 / 9.15 / 41.04 / 25.52	17.72 / 7.44 / 34.34 / 19.44
A-Mem	27.02 / 12.14 / 44.65 / 45.85	20.09 / 12.00 / 37.06 / 36.67
A-Mem*	20.76 / 9.22 / 33.34 / 35.40	14.90 / 8.81 / 27.58 / 31.08
LangMem	35.51 / 26.04 / 40.91 / 30.75	26.86 / 22.32 / 33.63 / 25.84
Zep	35.74 / 19.37 / 49.56 / 42.00	23.30 / 14.82 / 38.92 / 34.53
OpenAI	34.30 / 20.09 / 39.31 / 14.04	23.72 / 15.42 / 31.16 / 11.25
Mem0	38.72 / 28.64 / 47.65 / 48.93	27.13 / 21.58 / 38.72 / 40.51
$Mem0^g$	38.09 / 24.32 / 49.27 / 51.55	26.03 / 18.82 / 40.30 / 40.28