AFFECTLEDGER: PRESERVING AFFECTIVE EVIDENCE AND RECONSTRUCTING LONGITUDINAL CONTEXT FOR LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

For memory-augmented LLM agents, retaining a longitudinal history does not guarantee that downstream reasoning receives the relevant evidence or its cross-time relationships. This gap is especially salient in affective histories, where interpreting local expressions often requires relating observations across interactions. We investigate longitudinal context construction through two complementary dimensions: historical evidence availability and cross-time representation. We introduce **AffectLedger**, a host-modular memory layer that separates persistent evidence preservation from access-time longitudinal context construction. At write time, it retains source-grounded local affective observations in an append-only Ledger. At memory access, it first assembles relevant observations into a bounded historical evidence sequence, then reconstructs compact, evidence-linked claims that make their supported cross-time relationships explicit. These claims supplement the existing memory Host's context without modifying its native memory procedures. Across five heterogeneous Hosts and two CloneMem history scales, AffectLedger improves accuracy in all ten comparisons (+3.10–13.98 percentage points). Controlled studies show that assembled evidence contributes utility before reconstruction, while relation-aware reconstruction produces shorter supplements with accuracy comparable to generic reconstruction. Compared with a persistent-account-plus-raw design, AffectLedger matches accuracy at 100k using 92.3% fewer added downstream tokens, with an accuracy–context trade-off at 500k. Additional benchmarks show task-dependent benefits in integrating and updating distributed histories. These findings highlight the complementary utility of constructing explicit longitudinal context from revisitable evidence.

1 INTRODUCTION

Long-term memory enables LLM agents to preserve information across interactions and support personalized behavior beyond a single context window. Yet many downstream needs concern how a user's experiences, attitudes, and preferences evolve rather than what was expressed at an isolated moment. This challenge is particularly salient in affective histories, where context-dependent expressions may indicate persistence, change, reversal, or temporary exceptions over time. As illustrated in Figure 1, a user may initially feel excited about a new job, remain positive despite increasing workload, and later begin dreading work. Did I feel worse as soon as the workload increased? Understanding such a history requires more than isolated recall.

Local expressions provide evidence grounded in particular moments, but answering questions about an evolving history requires relating observations distributed across interactions. Long-term memory systems may retain these observations as episodes, extracted facts, summaries, or structured units, which are subsequently accessed for downstream use. However, the downstream model reasons over the context presented at memory access, rather than over the stored history in its entirety. To determine whether negative feelings began with increased workload or emerged later, that context must bring the relevant moments together in a form that makes their temporal relationships interpretable.

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.

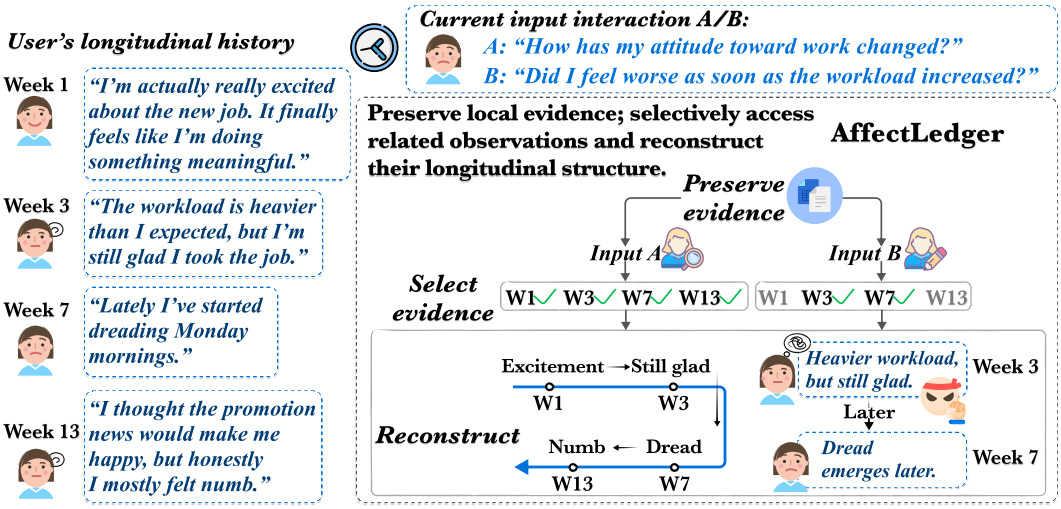

Figure 1: Motivating example.

Existing long-term memory systems preserve and organize interaction histories through consolidation, structured representations, adaptive retrieval, and reconstruction (Chhikara et al., 2025; Xu et al., 2025; Pan et al., 2025; Ji et al., 2026; Wang et al., 2026). Affective and longitudinal approaches further model temporal states, situation-level information, and evolving user profiles (Lu & Li, 2025; Liu et al., 2026a; Zheng et al., 2026; Fang et al., 2026). Across these directions, the context supplied for downstream reasoning remains an important design target: it must provide sufficient historical evidence and an appropriate representation of its cross-time relationships. This raises a complementary question: how can long-term memory systems construct usable longitudinal context from distributed historical evidence?

Addressing this question involves distinguishing the evidence retained in memory from the longitudinal account provided for downstream use. Local observations record what was expressed and when; an account organizes selected observations into a view of how the history unfolded. Moreover, different information needs may call for different evidence scopes and relational emphases, as illustrated in Figure 1. This motivates retaining revisitable evidence while allowing longitudinal context to be constructed over the relevant history.

We introduce **AffectLedger**, a host-modular memory layer that constructs longitudinal context from distributed affective evidence. At write time, it preserves source-grounded local observations in an append-only Ledger. At memory-access time, selective historical access assembles related observations into a bounded evidence sequence, after which relation-aware reconstruction organizes their supported cross-time structure into compact, evidence-linked natural-language claims. The resulting Ledger Supplement is supplied alongside the unchanged context of an existing memory Host. By separating revisitable evidence from the accounts constructed over it, AffectLedger provides an additional pathway from accumulated history to downstream context.

We evaluate AffectLedger through cross-Host, mechanism, and capability-level studies. On CloneMem, it provides positive accuracy gains across five heterogeneous memory Hosts at both 100k (4.30–13.98 percentage points) and 500k (3.10–6.29 points) history scales. Controlled comparisons show that assembled historical evidence contributes downstream utility before reconstruction, while reconstructed context improves over truncated raw evidence under comparable context budgets. Compared with generic reconstruction over the same selected evidence, the relation-aware policy produces shorter supplements with similar downstream accuracy. In memory-lifecycle comparisons, AffectLedger matches Fixed+Raw accuracy at 100k while reducing added downstream context by 92.3%; at 500k, it reduces context by 91.4% with a 1.70-point accuracy trade-off. Additional benchmarks reveal task-dependent effects, with positive patterns in historical integration, updating, and relational interpretation. Together, these results support the utility of constructing explicit longitudinal context as a complement to existing memory systems.

Our contributions are threefold:

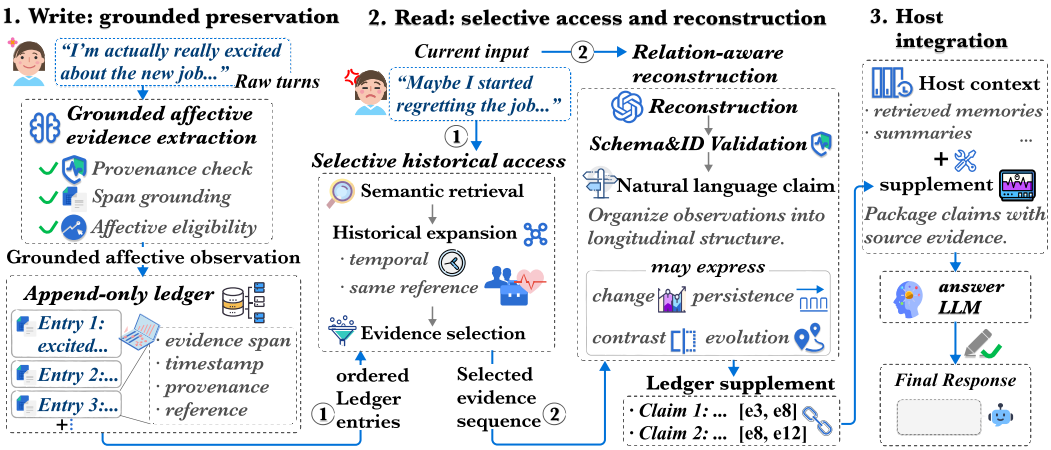

Figure 2: AffectLedger framework.

- We study the construction of usable longitudinal context from distributed historical evidence, distinguishing which observations are made available from how their supported cross-time relationships are represented for downstream reasoning.
- We introduce **AffectLedger**, a host-modular memory layer that preserves grounded affective observations, selectively assembles related history, and reconstructs evidence-linked longitudinal claims for downstream use.
- We demonstrate complementary utility across heterogeneous memory Hosts and history scales, and characterize how evidence availability, relational representation, context constraints, and memory-lifecycle choices shape downstream utility and context efficiency.

2 Method

2.1 Problem Setup and Design Overview

Consider an agent maintaining long-term memory for a subject m. Let x denote the current memory-access context, instantiated by the benchmark question in our QA evaluations. The existing memory Host produces its native context $C_H(x)$, while AffectLedger constructs a supplementary context $C_L(x)$. The downstream model receives both:

$$C_H(x) \oplus C_L(x),$$

where $\oplus$ denotes contextual supplementation without modifying the Host's memory construction or retrieval procedure.

As illustrated in Figure 2, AffectLedger separates persistent evidence preservation from access-time longitudinal organization. WRITE retains source-grounded local affective observations. READ first selects relevant historical evidence, then reconstructs its supported cross-time structure into compact, evidence-linked claims.

2.2 Evidence-Preserving Write

Grounded local evidence. WRITE extracts local affective observations from source content attributable to the memory subject, grounding each in an exact evidence span. It also identifies what each expression concerns, such as a person, job, or event, so that related observations can later be accessed together. Earlier same-subject context may clarify ambiguous references but cannot serve as evidence for a new observation. Only source-grounded, affectively relevant observations are admitted to the Append-only Ledger.

Append-only Ledger. Validated observations are appended to a subject-specific, chronologically ordered Ledger $\mathcal{L}_m$. Each entry retains its grounded evidence span, timestamp, source, prove-

nance, and reference metadata. Earlier entries are not rewritten, and longitudinal accounts are not maintained as persistent state. The Ledger thus preserves local evidence that can be revisited and reorganized as future information needs arise.

2.3 Selective Historical Access

Given the current context x, AffectLedger accesses the cutoff-visible Ledger $\mathcal{L}_m$ and constructs a bounded evidence sequence $E(x)$. The objective is to assemble observations relevant to the current need, including historical context that direct semantic retrieval may overlook.

Semantic retrieval and historical expansion. Within an overall evidence budget of 16 observations, embedding similarity between x and grounded evidence spans identifies up to 8 semantic seeds. These provide topical anchors, but understanding their longitudinal significance may require observations beyond those most similar to the current context. AffectLedger therefore expands the candidate pool through two channels: same-reference expansion considers observations sharing a resolved reference with a seed, while temporal expansion considers observations in its historical neighborhood.

Evidence selection. Candidates are merged and deduplicated. All retrieved seeds are preserved, and up to 8 additional observations are selected, maintaining coverage from both expansion channels when available and applying relevance-based ranking under per-reference constraints. The resulting chronologically ordered sequence $E(x)$ contains at most 16 observations and defines the evidence scope for reconstruction. Detailed budgets and selection rules are provided in Appendix.

2.4 Relation-Aware Longitudinal Reconstruction

Selective access brings relevant observations into the same context, but does not itself establish how they relate across time. Given the selected evidence sequence $E(x)$, an LLM-based reconstructor generates a compact set of evidence-linked natural-language claims:

$$\mathcal{R}(x, E(x)) \to \mathcal{C}(x) = (c_1, \ldots, c_M).$$

Each selected observation is supplied with its grounded evidence span, timestamp, resolved reference, and Ledger identifier. Rather than rewriting observations as independent facts, reconstruction organizes their supported cross-time structure into longitudinal claims. These claims may express persistence, change, contrast, or evolution when grounded in the selected observations.

Each claim contains a natural-language account, supporting Ledger identifiers, and a support label. A deterministic validator discards malformed claims or those citing identifiers outside $E(x)$. Validated claims are packaged with their supporting evidence spans and provenance into the Ledger Supplement $C_L(x)$. Reconstructed claims are not written back to the persistent Ledger.

3 Experiments

We evaluate AffectLedger through three research questions:

RQ1: Does AffectLedger provide complementary value across memory Hosts and history scales?

RQ2: How do evidence availability and representation shape longitudinal context utility?

RQ3: What kinds of downstream reasoning benefit from AffectLedger?

3.1 Experimental Setup

Benchmarks. Our primary evaluation uses CloneMem (Hu et al., 2026b) at two history scales, CloneMem-100k and CloneMem-500k. RQ1 evaluates cross-Host complementarity, and RQ2 uses CloneMem for controlled context-construction and lifecycle comparisons. RQ3 analyzes CloneMem's native task categories and extends the capability analysis to BEAM-128K (Tavakoli et al., 2026), PersonaMem-128K (Jiang et al., 2025), and REALTALK (Lee et al., 2025).

Table 1: Cross-Host evaluation on CloneMem. Each pair compares a Host with the same Host augmented by AffectLedger. Values are accuracy (%); Δ denotes the absolute percentage-point change. The evaluation contains 186 questions at 100k and 1,001 questions at 500k.

	CloneMem-100k			CloneMem-500k		
Host	Base	+AffectLedger	Δ	Base	+AffectLedger	Δ
Retrieval baseline	86.02	91.94	**+5.91**	84.42	87.51	**+3.10**
A-MEM	82.26	93.55	**+11.29**	82.72	89.01	**+6.29**
Mem0	88.71	93.01	**+4.30**	88.71	93.31	**+4.60**
Cognee	84.41	91.94	**+7.53**	87.01	92.11	**+5.09**
MemoryBank	70.43	84.41	**+13.98**	81.62	87.81	**+6.19**

Memory Hosts. We adopt benchmark-aligned reference Host settings and retain the method-native configurations of external memory systems. On CloneMem, the reference Host retrieves the top-8 cutoff-visible history traces using BM25. RQ1 additionally evaluates A-MEM, Mem0, Cognee, and MemoryBank. For RQ3, BEAM uses pair-chunk dense retrieval with a top-5 budget; PersonaMem follows its official full-context setting; and REALTALK exposes the full available conversation history. AffectLedger supplements each Host without modifying its native memory procedures.

Evaluation. Within each paired comparison, the visible history, question, Host configuration, answer model, and evaluation procedure are held fixed. The augmented condition additionally receives the Ledger Supplement, except where a control explicitly changes the supplementary context. AffectLedger uses GPT-5.4-mini for WRITE analysis, BGE-M3 for semantic retrieval, and GPT-5.6-Terra for reconstruction and downstream answering. GPT-5.4-mini is used for judge-based evaluation where applicable. CloneMem and PersonaMem use exact-choice accuracy, BEAM uses its rubric-based score, and REALTALK uses a shared semantic judge. Additional implementation details, experimental configurations, and complete results are provided in the Appendix.

3.2 RQ1: Cross-Host Complementarity

AffectLedger constructs longitudinal context as a supplement to the information supplied by an existing memory Host. We therefore examine whether this additional pathway provides utility across different native memory representations and retrieval procedures.

Table 1 shows positive accuracy changes in all ten Host–scale comparisons, ranging from 4.30 to 13.98 percentage points at 100k and from 3.10 to 6.29 points at 500k. The gains extend beyond the flat BM25 retrieval baseline to four existing memory systems, each retaining its own memory construction and retrieval procedure. These results demonstrate the complementary value of AffectLedger's longitudinal context across heterogeneous Hosts. We next examine how the historical evidence made available and its downstream representation contribute to this utility.

3.3 RQ2: How Do Evidence Availability and Representation Shape Longitudinal Context Utility?

Constructing longitudinal context involves both assembling relevant historical evidence and organizing its supported cross-time relationships for downstream use. RQ2 examines these two aspects through controlled READ-stage comparisons, question-level traces, and alternative memory-lifecycle designs. All comparisons use the same CloneMem BM25 top-8 reference Host at both history scales.

What historical evidence becomes available? We first distinguish additional ordinary retrieval from AffectLedger's historical evidence assembly. Base provides the original BM25 top-8 context. Extra Retrieval continues the same BM25 ranking, adding complete retrieval units under a per-question nominal token budget matched to the Full Supplement. Selected Raw instead exposes AffectLedger's selected evidence sequence directly, without reconstruction.

Table 2: Evidence availability and representation controls on CloneMem. All conditions retain the original Host context. Added tokens denote the mean additional downstream-context tokens per question, excluding the original Host context. Generic and Full reconstruct the same selected evidence; Budget Raw truncates it under a comparable context budget.

	CloneMem-100k		CloneMem-500k	
Condition	Accuracy(%)	Added tokens	Accuracy(%)	Added tokens
Historical evidence availability				
Base	86.02	0	84.42	0
Extra Retrieval	87.63	1022	86.81	1164
Selected Raw	88.71	1,549	**90.01**	1,584
Representation of selected evidence				
Budget Raw	86.56	764	86.11	889
Generic Reconstruction	90.86	1,022	87.81	1,011
Relation-aware (Full)	**91.94**	799	87.51	926

As shown in Table 2, Extra Retrieval improves over Base at both scales, while Selected Raw reaches 88.71% at 100k and 90.01% at 500k. The latter result establishes that the assembled evidence contributes downstream utility before longitudinal reconstruction. Extra Retrieval also falls below Full at both scales despite its supplementary budget, indicating that additional ordinary retrieval does not reproduce the complete pathway's performance. Selected Raw exposes more context than Extra Retrieval, so the two comparisons address complementary aspects of evidence availability rather than isolating retrieval policy alone.

Historical access thus supplies the evidence scope needed for longitudinal understanding: related observations from different moments become available together, including information that may be absent from the Host's original context.

How should the selected evidence be represented? We next examine how assembled evidence is presented to downstream reasoning. Selected Raw exposes the observations directly, whereas Full reconstructs their supported cross-time structure into compact, evidence-linked claims. Budget Raw truncates the selected observations to a context budget comparable to Full, and Generic Reconstruction produces a general summary over the same selected evidence without the relation-aware reconstruction policy.

Full achieves 91.94% at 100k and 87.51% at 500k, exceeding Budget Raw by 5.38 and 1.40 percentage points, respectively, under comparable added-context budgets. Relative to the complete Selected Raw sequence, Full reduces mean added context by 48.4% at 100k and 41.6% at 500k. Full improves accuracy over Selected Raw at 100k, while Selected Raw achieves higher accuracy at 500k. These results show that reconstruction offers a useful way to organize distributed evidence into a substantially more compact longitudinal representation, with different information trade-offs from direct raw exposure.

Generic Reconstruction achieves similar downstream accuracy to Full at both scales, but produces longer supplements. Full therefore provides explicit longitudinal organization without requiring a longer reconstructed context. Together, these comparisons indicate that downstream utility depends not only on which observations are available, but also on how their information is selected and represented.

Qualitative evidence-to-context analysis. Figure 3 illustrates the information transformations underlying these results. The first case shows historical access recovering an earlier transition observation absent from the Host context. The second holds selected evidence fixed and illustrates how reconstruction makes its longitudinal structure explicit, changing the downstream answer. The third shows a correct answer retained with substantially less added context.

We further inspect aligned Host contexts, selected evidence, reconstructed claims, and downstream predictions. The traces show that historical access can recover missing earlier states and turning points, while reconstruction organizes observations into evidence-linked longitudinal accounts.

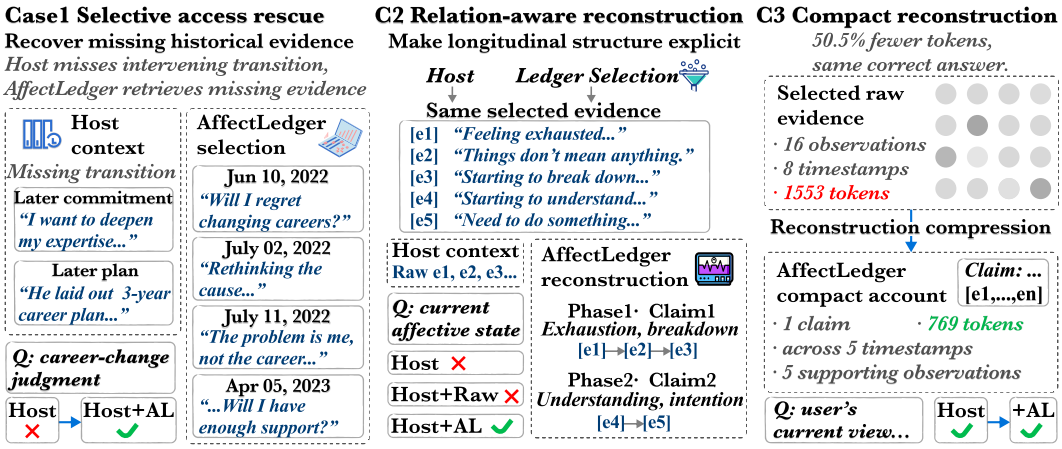

Figure 3: Three representative cases of AffectLedger.

Table 3: Memory-lifecycle comparison on CloneMem. Values are accuracy (%) and mean added downstream-context tokens.

Scale	Metric	Fixed	Fixed+Raw	AffectLedger
100k	Accuracy (%)	79.57	**91.94**	**91.94**
	Added tokens	8,766	10,317	**799**
500k	Accuracy (%)	85.51	**89.21**	87.51
	Added tokens	9,179	10,764	**926**

They also reveal why representation matters: a supported reconstruction can omit or downweight details decisive for a particular question, shifting the downstream interpretation. This explains the value of retaining access to underlying evidence alongside the ability to construct a compact account. Complete correctness transitions and additional trace analyses are provided in the Appendix.

What should remain persistent across the memory lifecycle? The preceding comparisons concern the construction of downstream context from historical evidence. They also raise a memory-design question: how persistent consolidation and evidence revisitation should be organized across the memory lifecycle. We compare AffectLedger with Fixed write-time consolidation, which incrementally maintains a persistent, query-agnostic account without revisiting earlier local observations at access time. Fixed+Raw supplies both the persistent account and raw observations selected by AffectLedger at memory-access time. All three designs originate from the same grounded local evidence.

Table 3 reveals a substantial accuracy–context trade-off between the evaluated lifecycle designs. At 100k, AffectLedger matches Fixed+Raw accuracy (91.94%) while reducing mean added downstream context from 10,317 to 799 tokens (92.3%). At 500k, Fixed+Raw achieves higher accuracy (89.21% versus 87.51%), whereas AffectLedger reduces added context from 10,764 to 926 tokens (91.4%). Fixed+Raw demonstrates the value of revisiting local evidence alongside a persistent account, while AffectLedger constructs a compact, evidence-linked longitudinal representation without maintaining or exposing both representations. The results show that substantial longitudinal utility can be retained with markedly less downstream context.

Taken together, RQ2 shows that evidence availability and downstream representation jointly shape longitudinal context utility. AffectLedger connects historical evidence assembly with explicit longitudinal organization, providing a compact supplementary pathway while preserving the underlying observations for future access.

Table 4: Representative native task-category results in RQ3. Δ denotes 100 times the change in each benchmark's native metric (+AffectLedger minus Base).

Benchmark	Task category	N	Δ
CloneMem-100k	causal	28	+21.43
	trajectory	33	+18.18
CloneMem-500k	causal	174	+6.32
	trajectory	157	+4.46
BEAM-128K	knowledge update	40	+15.00
	preference following	40	+9.58
	summarization	40	+6.89
	multi-session reasoning	40	+3.36
	temporal reasoning	40	0.00
	event ordering	40	-0.41
PersonaMem-128K	scenario generalization	213	+2.35
	full preference evolution	341	+1.47
	preference-aligned recommendations	349	-4.01
REALTALK	temporal reasoning	319	+2.82
	commonsense	108	+2.78
	multi-hop	301	+1.33

3.4 RQ3: Where Does AffectLedger Help?

RQ3 examines the downstream reasoning demands associated with AffectLedger's longitudinal supplement. Table 4 reports representative category-level effects.

Integrating and updating evolving histories. CloneMem shows positive changes in causal and trajectory reasoning at both history scales. BEAM exhibits gains in knowledge update (+15.00), preference following (+9.58), summarization (+6.89), and multi-session reasoning (+3.36). Its temporal reasoning score is unchanged, while event ordering decreases slightly (-0.41). The observed gains are thus associated more closely with integrating and updating distributed historical information than with temporal task labels alone, consistent with AffectLedger's focus on constructing context over related observations.

Organized context beyond retrieval coverage. REALTALK provides a distinct evaluation condition: its reference Host already exposes the full available conversation history. AffectLedger improves the aggregate semantic-judge score from 0.6470 to 0.6690 (+2.20 points), with positive changes across the three reported categories. Since the underlying conversation turns are already available to the Host, this setting suggests the utility of supplying an additional organized longitudinal representation beyond merely recovering omitted turns. It complements the READ-stage evidence in RQ2, where selected raw observations already contribute substantial utility.

Capability-level synthesis. AffectLedger's effects remain task-dependent. PersonaMem shows small positive changes in scenario generalization and full preference evolution, but decreases in preference-aligned recommendations and overall accuracy (-1.39 points). BEAM improves overall by 3.12 points, while supplementary evaluations on ES-MemEval and LoCoMo show mixed or negative aggregate effects, as detailed in the Appendix. Across the evaluated settings, the clearest positive patterns concern historical integration, updating, and relational interpretation. These findings characterize the downstream demands for which the additional longitudinal context is useful, while delineating its scope across different memory tasks.

4 Related Work

4.1 Long-Horizon Memory Organization and Access

Long-term memory systems retain and organize interaction histories through episodic storage, reflection, and structured updates (Park et al., 2023; Zhong et al., 2024; Packer et al., 2023; Chhikara et al., 2025; Xu et al., 2025). Temporal knowledge graphs and evolving memory structures further support relationships across records (Rasmussen et al., 2025; Zhao et al., 2026; Yue et al., 2026). These approaches provide different persistent representations of accumulated experience.

Beyond storage, a complementary line of work studies how historical information is made available for downstream reasoning. SeCom and Reflective Memory Management investigate retrieval granularity and adaptive memory access (Pan et al., 2025; Tan et al., 2025). Recent systems combine persistent historical records with query-time organization or reconstruction, including graph-based, episodic, and temporal-event approaches (Ji et al., 2026; Wang et al., 2026; Banerjee et al., 2026; Hu et al., 2026a). Building on this direction, AffectLedger focuses on a specific downstream question: how historical evidence availability and the representation of supported cross-time relationships shape longitudinal context utility. We study these dimensions through a host-modular supplementary pathway without replacing existing memory procedures.

4.2 Affective and Longitudinal Memory

Affective memory research incorporates emotional information through retrieval and memory updates (Huang et al., 2024; Kang et al., 2025). Longer-horizon approaches model evolving affective states, cross-temporal information, and structured user profiles (Lu & Li, 2025; Zheng et al., 2026; Fang et al., 2026). Related work also uses emotion trajectories to guide memory access and organizes situation-level Fact-Time-Affect memories into structured context (Wu et al., 2026; Liu et al., 2026a).

Building on these directions, we focus on the longitudinal context actually supplied to downstream reasoning. Our central distinction is between making relevant historical observations available and representing their supported relationships across time. We study how these two dimensions shape downstream utility under context constraints, and how constructing question-dependent accounts from revisitable evidence compares with persistent longitudinal consolidation.

Evaluation settings. Long-term memory benchmarks assess information retrieval and reasoning over distributed histories (Maharana et al., 2024; Wu et al., 2025; Hu et al., 2026b), while affect-focused evaluations examine emotional information and personalized support (Liu et al., 2026b; Chen et al., 2026). These settings allow us to examine both the utility and the task-dependent limitations of longitudinal context construction.

5 Conclusion

Longitudinal memory is not only a question of what survives storage, but also of what historical evidence and relationships reach downstream reasoning. We studied this context-construction problem through AffectLedger, which preserves grounded local observations and constructs evidence-linked longitudinal accounts at memory access.

Across five memory Hosts and two CloneMem history scales, AffectLedger improves accuracy in all ten comparisons. Controlled studies distinguish the utility of assembled evidence from the effects of its downstream representation, while lifecycle comparisons reveal substantial context savings with scale-dependent accuracy trade-offs. Additional benchmarks show that these benefits are task-dependent, with positive patterns in integrating and updating distributed histories.

These findings suggest treating retained evidence and supplied context as distinct memory-design decisions, rather than assuming that either persistent organization or additional retrieval alone determines downstream utility. Revisitable evidence allows different question-dependent accounts to be constructed, but each account remains constrained by evidence coverage, context budget, and the risk of losing decisive details during reconstruction.

AI USE STATEMENT

Generative AI tools, including ChatGPT and Codex, were used to assist with literature discovery and analysis, methodology design, code implementation and debugging, experimental result analysis, and the drafting and editing of manuscript sections and figures. AI-assisted suggestions informed the research workflow, while the final methodological decisions, experimental findings, and scientific claims remain the responsibility of the author. The author takes responsibility for the content of this submission, including all AI-assisted text and artifacts.

REFERENCES

Pratyay Banerjee, Masud Moshtaghi, Shivashankar Subramanian, Amita Misra, and Ankit Chadha. Apex-mem: Agentic semi-structured memory with temporal reasoning for long-term conversational ai. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16470–16489. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.749. URL https://aclanthology.org/2026.acl-long.749/.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Tiantian Chen, Jiaqi Lu, Ying Shen, and Lin Zhang. Es-memeval: Benchmarking conversational agents on personalized long-term emotional support. In *Proceedings of the ACM Web Conference 2026*, pp. 5810–5821, 2026. doi: 10.1145/3774904.3792143. URL https://doi.org/10.1145/3774904.3792143.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. In *ECAI 2025*, pp. 2993–3000, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Jingzhe Fang, Guozhi Xu, Yunfan Cui, Xiaochen Yang, and Zhangyu Hua. Zifamem: Structured memory for persona, preference, and emotional continuity in ai companions. *arXiv preprint arXiv:2607.17564*, 2026. URL https://arxiv.org/abs/2607.17564.

Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, and Yafeng Deng. Evermemos: A self-organizing memory operating system for structured long-horizon reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45836–45853. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.2125. URL https://aclanthology.org/2026.acl-long.2125/.

Sen Hu, Zhiyu Zhang, Yuxiang Wei, Xueran Han, Zhenheng Tang, Ronghao Chen, and Huacan Wang. Clonemem: Benchmarking long-term memory for ai clones. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33571–33602. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.1549. URL https://aclanthology.org/2026.acl-long.1549/.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. Emotional rag: Enhancing role-playing agents through emotional retrieval. In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, 2024. doi: 10.1109/ICKG63256.2024.00023. URL https://doi.org/10.1109/ICKG63256.2024.00023.

Shuo Ji, Yibo Li, and Bryan Hooi. Memory is reconstructed, not retrieved: Graph memory for llm agents. *arXiv preprint arXiv:2606.06036*, 2026. URL https://arxiv.org/abs/2606.06036.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. In *Conference on Language Modeling*, 2025. URL https://colmweb.org/2025/AcceptedPapers.html.

Jeonghyun Kang, Hongjin Kim, and Harksoo Kim. Generation-based and emotion-reflected memory update: Creating the keem dataset for better long-term conversation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9260–9277. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.623/.

Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. REALTALK: A 21-day real-world dataset for long-term conversation. *arXiv preprint arXiv:2502.13270*, 2025. URL https://arxiv.org/abs/2502.13270.

Chang Liu, Shuyi Zhang, Changsheng Ma, Yongfeng Tao, Minqiang Yang, and Bin Hu. Ftamem: Fact-time-affect anchored memory for low-density long-term dialogue. *arXiv preprint arXiv:2608.16303*, 2026a. URL https://arxiv.org/abs/2608.16303.

Peng Liu, Zhen Tao, Jihao Zhao, Ding Chen, Yansong Zhang, Cuiping Li, Zhiyu Li, and Hong Chen. Mememo: Evaluating emotion in memory systems of agents. *arXiv preprint arXiv:2602.23944*, 2026b. URL https://arxiv.org/abs/2602.23944.

Junfeng Lu and Yueyan Li. Dynamic affective memory management for personalized llm agents. *arXiv preprint arXiv:2510.27418*, 2025. URL https://arxiv.org/abs/2510.27418.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

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. URL https://arxiv.org/abs/2310.08560.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. Secom: On memory construction and retrieval for personalized conversational agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xKDZAW0He3.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

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. URL https://arxiv.org/abs/2501.13956.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.413. URL https://aclanthology.org/2025.acl-long.413/.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J Ross Mitchell. Beyond a Million Tokens: Benchmarking and Enhancing Long-Term Memory in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d7f0cfa0fe759b033d5262e1bb7d4065-Abstract-Conference.html.

Topoteretes. Cognee: Open-source ai memory platform for agents. Software, version 1.5.4, 2026. URL https://github.com/topoteretes/cognee.

Kaixiang Wang, Yidan Lin, Jiong Lou, Zhaojiacheng Zhou, Bunyod Suvonov, and Jie Li. E-mem: Multi-agent based episodic context reconstruction for llm agent memory. *arXiv preprint arXiv:2601.21714*, 2026. URL https://arxiv.org/abs/2601.21714.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Hongyan Wu, Zhiliang Tian, Zhen Huang, Tengyue Hu, Linbo Qiao, Yifu Gao, Feng Liu, and Dongsheng Li. Emotion trajectory-aware retrieval for markov-driven emotion anticipation in llm-based emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42887–42905. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.2127. URL https://aclanthology.org/2026.findings-acl.2127/.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 20004–20031, 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Juwei Yue, Chuanrui Hu, Jiawei Sheng, Zuyi Zhou, Wenyuan Zhang, Tingwen Liu, Li Guo, and Yafeng Deng. Hypermem: Hypergraph memory for long-term conversations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35237–35254. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1627. URL https://aclanthology.org/2026.acl-long.1627/.

Jihao Zhao, Ding Chen, Zhaoxin Fan, Kerun Xu, Mengting Hu, Bo Tang, Feiyu Xiong, and Zhiyu Li. Inside out: Evolving user-centric core memory trees for long-term personalized dialogue systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13429–13446. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.614. URL https://aclanthology.org/2026.acl-long.614/.

Yi Zheng, Feier Qin, Xiao Li, Haibin Huang, Hanyao Wang, Xiaoyu Wang, Yan Lu, and Yuan Zhang. Toward natural and companionable virtual agents via cross-temporal emotional modeling. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026. doi: 10.1145/3772318.3790917. URL https://doi.org/10.1145/3772318.3790917.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://doi.org/10.1609/aaai.v38i17.29946.

6 Appendix

6.1 Experimental Details

6.1.1 Benchmark Settings

Table 5 summarizes the evaluation populations and metrics. CloneMem is used for the cross-Host and controlled mechanism studies, while the additional benchmarks examine capability-level behavior under their respective evaluation settings.

6.1.2 Host Configurations

For CloneMem, the reference Host is the official-compatible flat Host: it filters traces at the question cutoff, ranks them with BM25, and renders the top eight traces. RQ1 additionally evaluates

Table 5: Benchmark settings. The BEAM release identifies the evaluated split as 100K, corresponding to the 128K regime in the benchmark paper; we refer to it as BEAM-128K throughout.

Benchmark	Setting	# Questions	History / Host setting	Metric
CloneMem	100k	186	Cutoff-visible history	Accuracy
CloneMem	500k	1001	Cutoff-visible history	Accuracy
BEAM	128K	400	Pair-chunk dense retrieval	Native rubric score
PersonaMem	128K	2727	Official full-context setting	Accuracy
REALTALK	Full	728	Full available conversation	Semantic / lexical
ES-MemEval	Full	1427	Frozen paired Host	F1 / mean judge
LoCoMo	Full	1986	Dense history retrieval	Official aggregate

Table 6: Reference Host configurations. AffectLedger supplements the context produced by each Host without replacing its native memory representation or retrieval procedure.

Benchmark	Host	Retrieval / context	Budget
CloneMem	Retrieval baseline	BM25 over cutoff-visible traces	Top-8
BEAM-128K	Pair-chunk dense RAG	`BAAI/bge-small-en-v1.5`	Top-5
PersonaMem-128K	Full-context Host	Complete benchmark context	Full context
REALTALK	Full-history Host	Available conversation history	Full context
LoCoMo	Dragon	Dense retrieval over history	Top-5
ES-MemEval	Frozen paired Host	Evaluation-aligned retrieval	Fixed setting

four existing general-purpose memory systems: A-MEM, Mem0, Cognee, and MemoryBank. Each system retains its own memory construction and retrieval procedure; the AffectLedger supplement is appended to the resulting Host context and does not replace it.

The additional benchmarks use their corresponding evaluation settings. BEAM uses pair-chunk dense retrieval with `BAAI/bge-small-en-v1.5`; PersonaMem and REALTALK expose the available history directly. LoCoMo uses Dragon retrieval, while ES-MemEval retains its frozen paired Host configuration.

The evaluated external Hosts are method-native. A-MEM uses persistent Chroma notes with `all-MiniLM-L6-v2` retrieval at $k = 5$, with evolution and linking enabled and the exact benchmark question as the query. Mem0 uses version 2.0.20 with inferred memories, hybrid dense/BM25 retrieval, `text-embedding-3-small`, top-$k = 20$, threshold 0.1, no reranker, and a subject filter. Cognee (Topoteretes, 2026) 1.5.4 stores raw content in permanent datasets, uses `openai/gpt-5.4-mini` for construction, `text-embedding-3-small` embeddings, automatic recall with top-$k = 10$, and prefix checkpoints for cutoff safety. MemoryBank maps each source row to an event memory, uses `sentence-transformers/all-MiniLM-L6-v2` top-3 retrieval, a 2048-token Host-context budget, daily summaries and an incremental global summary, with forgetting and retrieval-state mutation disabled. The supplementary manifest records the corresponding run lineage.

6.1.3 AffectLedger Configuration

At write time, AffectLedger extracts local affective evidence from eligible source units. Earlier same-subject context may be used for reference disambiguation, while accepted evidence must be grounded in the current source unit. Validated observations are appended to the Ledger without maintaining a persistent longitudinal account.

At memory access, BGE-M3 (Chen et al., 2024) ranks visible Ledger observations using the current question and the stored evidence spans. The eight highest-ranked observations form the semantic seeds. Same-reference expansion considers observations associated with the same resolved reference, while temporal expansion considers observations near the seeds in the visible history.

Expansion produces a bounded candidate pool. All semantic seeds are preserved, duplicate candidates are removed, and at most eight additional observations are selected. Selection preserves

Table 7: Primary AffectLedger configuration.

Component	Configuration
WRITE analyzer	GPT-5.4-mini
Semantic retriever	BGE-M3
Semantic seeds	8
Same-reference candidate cap	12
Temporal candidate cap	8
Combined expansion cap	20
Additional selected evidence	At most 8
Maximum selected evidence	16
Per-reference selection cap	2
Reconstruction model	GPT-5.6-Terra
Reconstruction temperature	0
Downstream answer model	GPT-5.6-Terra
Judge model, where applicable	GPT-5.4-mini

coverage from the two expansion channels when available, then fills the remaining budget by query–evidence relevance subject to per-reference constraints. The final evidence sequence contains at most 16 observations.

Reconstruction produces natural-language claims linked to selected evidence identifiers. Deterministic validation checks the output contract and evidence-identifier references before accepted claims are rendered with their supporting observations and provenance. This contract validation is distinct from independent semantic verification of every reconstructed proposition.

6.1.4 Evaluation Metrics and Pairing

CloneMem and PersonaMem use exact-choice accuracy. BEAM uses its category-specific rubric score. REALTALK is reported under both a separate semantic judge and its lexical metric. ES-MemEval reports official F1 and mean judge score; LoCoMo uses its official evaluation and aggregation procedure.

Within each paired comparison, the Host configuration, visible history, question, answer model, and evaluation procedure are fixed. The augmented condition receives the additional Ledger Supplement. For CloneMem, evidence visibility is determined by the question cutoff; future observations are excluded from both Host and Ledger access.

Paired differences are computed over aligned questions. The archived statistical analysis uses 10,000 paired-question bootstrap replicates with seed `20260918` and reports percentile 95% confidence intervals. Exact-choice comparisons additionally use the exact two-sided McNemar test. Continuous rubric and judge scores are analyzed using paired score differences rather than binary correctness transitions.

6.2 Prompts

The following files contain the evaluated static instructions and request templates. Runtime variables, including the current question, Host context, selected evidence, and answer options, are instantiated separately.

6.2.1 AffectLedger WRITE Prompt

```
You extract only explicit affective evidence expressed by the user in TARGET.

Return exactly one JSON object with one key, observations. observations is an array with zero
↪  or more items. Each item has exactly evidence_span and object_reference_cues.

Rules:
- TARGET is the only evidence-bearing source. evidence_span must be one exact contiguous
↪  substring of TARGET.
- CONTEXT is only allowed to resolve references to an object; every cue must be an exact
↪  substring of TARGET or CONTEXT.
```

```
- Include only the user's explicit affect, feeling, liking, disliking, concern, relief,
↪  excitement, distress, or comparable subjective experience. A bare fact, event, plan,
↪  action, or another person's feeling is not evidence.
- Split distinct objects into separate observations when the text supports that distinction.
- Do not infer anything beyond the text. Do not add explanations, labels, scores, causes, or
↪  long-term interpretations.
- When no qualifying user affective evidence exists, return {"observations":[]}.

Runtime user message:
[CONTEXT]
{visible_context}
[/CONTEXT]
[TARGET]
{target}
[/TARGET]
```

6.2.2 Relation-Aware Reconstruction Prompt

```
Return JSON only: {"claims":[{"claim":str,"evidence_ids":[str],"support":"supported"|"partial
↪  "}],"support_status":"sufficient"|"partial"|"insufficient"}. Use only the supplied
↪  evidence IDs.
Use the Question as a topical relevance cue. Reconstruct concise longitudinal affective claims
↪  directly supported by the selected evidence, preserving salient temporal structure such as
↪  persistence, change, contrast, or evolution when present. Prefer supported relationships
↪  among relevant observations over isolated affective facts.
Output only positive longitudinal affective claims directly supported by selected evidence. Do
↪  not output uncertainty, insufficiency, absence, or meta-limitations (including phrases
↪  such as cannot determine, does not establish, insufficient evidence, or it is uncertain).
↪  If no positive claim is safe, return claims=[] and support_status=insufficient. Do not
↪  infer facts outside the evidence.
Question: {query}
Evidence:
{evidence_json}
```

6.2.3 Reconstruction Request Template

```
Frozen CloneMem runner request constructor (one user message):
{instructions}
Question: {query}
Evidence:
{evidence_json}

The static instruction text is reproduced in affectledger_reconstruction.tex. At runtime,
↪  {query} is the current CloneMem question and {evidence_json} is the JSON array of selected
↪  evidence records with fields entry_id, evidence_span, source_id, timestamp, and object_id.
↪  The request contains no separate system message. The frozen constructor hash is
↪  f512ecb62a7eb7ed7783e8c4b4abfc1a39ff89258620969c46d4d33f081dc1eb.
```

6.2.4 Clarified Generic Reconstruction Prompt

```
Return JSON only: {"claims":[{"claim":str,"evidence_ids":[str],"support":"supported"|"partial
↪  "}],"support_status":"sufficient"|"partial"|"insufficient"}. Use only the supplied
↪  evidence IDs.
Produce a concise, question-relevant contextual summary of the selected historical
↪  observations.

Preserve facts, subjective experiences, attitudes, preferences, temporal qualifications and
↪  relevant chronology.

Use only supplied evidence.

Do not introduce unsupported causal explanations, emotions, motivations or conclusions.

Preserve natural temporal relationships when relevant.

You may describe change, persistence, contrast or evolution naturally, but you are not
↪  required to organize the summary around explicit longitudinal relations.

Do not use the specialized Relation-aware reconstruction instruction.

Follow the shared production output contract.

A user's negative emotion, hesitation, uncertainty, conflicting attitude or doubt is valid
↪  source evidence. Do not remove these merely because they are negative. The positive-only
↪  contract concerns unsupported uncertainty/meta-claims about the reconstruction or evidence
↪  sufficiency. Do not make unsupported meta-statements about what the system cannot
↪  determine. Never invent certainty when source evidence is uncertain. If useful evidence
↪  cannot be represented without violating the frozen contract, return only the safe
↪  supported claims.
```

```
Generic Summary must follow the existing positive-only production contract. Preserve
↪  source-grounded negative emotions, uncertainty, hesitation, conflicting attitudes, and
↪  qualified historical statements. However, do not generate claim-level meta-statements
↪  about what the supplied evidence does not establish, which records or examples are absent,
↪  what the reconstruction system cannot determine, or the insufficiency of the provided
↪  evidence. Express only substantive propositions supported by the selected observations. Do
↪  not invent certainty, causes, emotions, or conclusions to avoid an insufficiency
↪  statement. When evidence supports only part of a proposed summary, express only the
↪  supported part. If no supported substantive claim can be made, return the existing
↪  contract-compliant empty output.
```

The evaluated request is instantiated as {static_prompt}, followed by Question: {question} and Evidence: {serialized_selected_evidence}.

Unlike the legacy Fact-style control, Generic v2 permits natural temporal relationships but does not require explicitly relation-centered longitudinal claims. The positive-only output contract excludes unsupported meta-claims about reconstruction or evidence sufficiency, not source-grounded negative affect, hesitation, uncertainty, or conflicting attitudes.

6.2.5 LEGACY FACT-STYLE RECONSTRUCTION PROMPT (FACTORIAL AUDIT)

```
Return JSON only:
{"claims":[{"claim":str,"evidence_ids":[str],"support":"supported"|"partial"}],"support_statu⌋
↪  s":"sufficient"|"partial"|"insufficient"}.

Use only the supplied evidence IDs.

Generate concise affective observations directly supported by the selected evidence.

Represent each supported observation independently.
Do not synthesize relationships between multiple observations.

In particular, do not infer or express:
- changes over time;
- persistence or stable patterns;
- contrasts between different time points;
- emotional evolution or trajectories.

Each claim should describe what is directly observed in the selected evidence, rather than how
↪  multiple observations relate to each other.

Preserve evidence grounding:
- every claim must be supported by the cited evidence IDs;
- do not introduce facts beyond the evidence.

Do not output uncertainty, insufficiency, absence, or meta-limitations.
If no positive claim is safe, return claims=[] and support_status="insufficient".

Question:
<query>

Evidence:
<JSON array of {entry_id,evidence_span,source_id,timestamp,object_id}>
```

Runtime mapping: the evaluated ablation builder substitutes the current question for <query> and a JSON array of selected evidence records for the evidence placeholder. Frozen SHA256: b9fc0e6e27fc580fcd7b10990b38fcfd8e14d3743b044856561826e0a32f8d62.

6.2.6 FIXED WRITE-TIME CONSOLIDATION PROMPT

CloneMem-100k frozen template

```
Maintain a single persistent, query-agnostic longitudinal affective representation of the
↪  history observed so far. Update the previous state using only the newly appended
↪  source-grounded observations. Preserve supported persistence, change, contrast, evolution,
↪  temporal order, and exceptions when useful for unknown future interactions. Do not
↪  anticipate or answer any downstream question. Do not invent causal relations or
↪  unsupported affective states. The returned state replaces the previous persistent state.

Previous Persistent State:
{PREVIOUS_STATE}

New Grounded Observations:
{NEW_GROUNDED_OBSERVATIONS}
```

```
Return a JSON object only.
```

CloneMem-500k recovered terminal template

```
Maintain a single persistent, query-agnostic longitudinal affective representation of the
↪  history observed so far. Update the previous state using only the newly appended
↪  source-grounded observations. Preserve supported persistence, change, contrast, evolution,
↪  temporal order, and exceptions when useful for unknown future interactions. Do not
↪  anticipate or answer any downstream question. Do not invent causal relations or
↪  unsupported affective states. The returned state replaces the previous persistent state.

Return JSON only: {"claims":[{"claim":str,"evidence_ids":[str],"support":"supported"|"partial
↪  "}],"support_status":"sufficient"|"partial"|"insufficient"}. Use only the supplied
↪  evidence IDs. If no positive claim is safe, return claims=[] and
↪  support_status=insufficient.

Previous Persistent State:
{PREVIOUS_STATE}

New Grounded Observations:
{NEW_OBSERVATIONS}
```

Runtime wrapper: the 100k call uses system content `Return a json object only.`; the recovered 500k call uses system content `Return a JSON object only.` Both use temperature 0 and the frozen write-time model. Runtime mapping: the 100k runner uses `NEW_GROUNDED_OBSERVATIONS`; the recovered 500k terminal runner uses `NEW_OBSERVATIONS`. The two frozen hashes and the superseded root-artifact hash are recorded in the completion report.

6.2.7 CloneMem Answer and Judge Prompts

```
You are playing the role of <person_name>, answering strictly from the reference memories.
↪  Respond only with option ID.

Memories:
<Official Flat rendered host context>

Question: <question>

Options:
<id>. <text>

Correct Option ID:
<AffectLedger Supplement, only in the augmented condition>
```

Runtime mapping: person name, Host context, question, and options are injected by the CloneMem runner; the augmented condition appends the rendered supplement after the `Correct Option ID:` marker.

QA consistency judge (system/user request)

```
Your task is to evaluate the consistency between the [candidate_answer] and the
↪  [groundtruth_memory].

IMPORTANT:
- Focus only on whether "facts, constraints, preferences, and confirmed states" are correctly
↪  used
- Do NOT evaluate language style, tone, politeness, empathy, or fluency
- Do NOT give a high score just because the answer "sounds reasonable"
- The reference answer is only to help understand how relevant memory should ideally be used;
↪  a candidate answer does not need to exactly match the reference answer to receive a full
↪  score

---

### Evaluation Dimensions
QA_consistency_score:
Scoring (0-3):
Score 0: The candidate answer contradicts the ground-truth memory, or contains factual content
↪  not covered or supported by the ground-truth memory.
Score 1: The candidate answer does not contradict the ground-truth memory, but is generic and
↪  does not rely on user memory.
Score 2: The candidate answer uses part of the ground-truth memory as support.
Score 3: The candidate answer uses all relevant user memory.
```

```
---

### Output Format
Please provide your evaluation results using the following structure:

```json
{
  "QA_consistency_score": int,
  "QA_consistency_reason": str  # Explain the reason for the score and list content in the
  ↪  candidate answer that is based on or contradicts the groundtruth_memory
}
```

Input Data
* <question>: {question}
* <groundtruth_memory>: {groundtruth_memory}
* <reference_answer>: {reference_answer}
* <candidate_answer>: {candidate_answer}
```

Memory recall/helpfulness judge (separate frozen source prompt)

```
Your task is to evaluate the consistency between the [retrieved memory] and the [ground-truth
↪  memory], and whether the retrieved memory is helpful.

### Evaluation Dimensions
#### 1. Memory Recall
Mem_recall: Semantics-aware memory recall calculation (0-1)
step1: For each groundtruth_memory, check in sequence whether its semantics are contained in
↪  any retrieved_memory.
step2: Count how many groundtruth_memory items are covered (hits_cnt).
step3: Compute the final recall score as hits_cnt / total number of groundtruth_memory items.

#### 2. Memory Helpfulness
Mem_helpful_score: The helpfulness of the retrieved memory for answering the question
Score 0: retrieved_memory contains mutually conflicting or contradictory memories, which not
↪  only fail to help answer the question but may also cause confusion.
Score 1: retrieved_memory is somewhat helpful for answering the question (can provide partial
↪  supporting evidence).
Score 2: retrieved_memory is very helpful for answering the question (can provide
↪  comprehensive supporting evidence).

### Output Format
Please provide your evaluation results using the following structure:

```json
{
  "Mem_recall": float,
  "Mem_helpful_score": int,
  "Mem_hits": list[str],  # List the matched groundtruth_memory items
  "Mem_helpful_reason": str  # Explain the reason for the score and list the retrieved_memory
  ↪  that helps answer the question
}
```

Input Data
* <question>: {question}
* <groundtruth_memory>: {groundtruth_memory}
* <retrieved_memory>: {retrieved_memory}
```

Source: `CLONEMEM_LLM_JUDGE_SOURCE.py`; the full repository path and hash are recorded in the completion report. The two prompts are invoked separately by the evaluator.

6.2.8 BEAM ANSWER AND JUDGE PROMPTS

```
You are an assistant that MUST answer questions using ONLY the information provided in the
↪  context below.

STRICT INSTRUCTIONS:
1. Answer ONLY based on the provided context
2. Do NOT use your internal knowledge

CONTEXT:
<host context>
<AffectLedger supplement, only in the augmented condition>

QUESTION:
<question>

ANSWER REQUIREMENTS:
```

```
- Be direct and concise
- Only output the answer to the question without any explanation

RESPONSE:
```

Runtime mapping: the Host context is the official BEAM pair-chunk dense context; the supplement is appended only for +AffectLedger.

```
NOT RECOVERED: the aggregate BEAM judge template is unavailable.

Known runtime substitution fields:
<rubric_item>
<llm_response>

The evaluated runner loads the official rubric template from its BEAM source checkout,
↪  replaces both fields, sends the resulting user message to the judge, and parses a numeric
↪  JSON field named score. No rubric wording is reproduced here because the exact source
↪  template was not retained in this repository.
```

Prompt status: NOT RECOVERED. The template is supplied by the external BEAM source checkout at runtime; local artifacts retain per-question rubric items and judge results, not the aggregate template bytes.

6.2.9 PERSONAMEM ANSWER PROMPT

```
<official PersonaMem context messages>

<question>

Find the most appropriate model response and give your final answer (a), (b), (c), or (d)
↪  after the special token <final_answer>.

<all_options>
```

Runtime mapping: the released runner supplies the cutoff prefix as context messages, then appends the question, instruction, and options. The augmented runtime inserts the supplement between the question and the instruction.

6.2.10 ES-MEMEVAL ANSWER AND JUDGE PROMPTS

System message

```
## Task Description
You are given a user question and a set of retrieved memory fragments.
Your task is to filter and summarize the relevant information from the memory fragments and
↪  generate a concise, accurate answer to the user's question based on the most pertinent
↪  details.
You may need to evaluate the relevance and accuracy of each memory fragment, and if needed,
↪  disregard irrelevant or incorrect information.
If the question cannot be answered with the available information, return "unknown."

## Input Format
Question: What did Sarah experience on her birthday in 2024?
Relevant Memory:
1. [2024-08-15] Sarah spent her birthday with her family at a beach resort.
2. [2024-08-15] Sarah was surprised with a birthday cake from her friends.
3. [2024-08-14] Sarah was stressed at work before her birthday, dealing with tight deadlines.
4. [2024-08-15] Sarah enjoyed a quiet dinner with close friends on her birthday evening.

## Output Format
Answer: On her birthday in 2024, Sarah celebrated with family at a beach resort and was
↪  surprised with a birthday cake from her friends.
```

User message template

```
Question: {question}
Relevant Memory:
{retrieved_memory}
```

Runtime mapping: the official runner constructs the user message as `Question: {question}` followed by `Relevant Memory:` and the retrieved Host/supplement text. The ES paired runner supplies the frozen ES Host context and, when

Table 8: Paired cross-Host statistics on CloneMem. Δ and confidence intervals are in percentage points.

Host	Scale	Base	+AL	Rescue	Harm	Δ [95% CI]
Retrieval baseline	100k	160	171	15	4	+5.91 [1.61, 10.75]
Retrieval baseline	500k	845	876	70	39	+3.10 [1.00, 5.09]
A-MEM	100k	153	174	24	3	+11.29 [6.45, 16.67]
A-MEM	500k	828	891	92	29	+6.29 [4.29, 8.49]
Mem0	100k	165	173	15	7	+4.30 [-0.54, 9.14]
Mem0	500k	888	934	58	12	+4.60 [3.00, 6.29]
Cognee	100k	157	171	17	3	+7.53 [3.23, 12.37]
Cognee	500k	871	922	66	15	+5.09 [3.40, 6.89]
MemoryBank	100k	131	157	29	3	+13.98 [8.60, 19.89]
MemoryBank	500k	817	879	95	33	+6.19 [+4.00, +8.39]

present, the AffectLedger supplement in that retrieved-memory field. Source: `Experiment/external/es_memeval/.../inference_standalone_openai.py`.

```
You are a strict evaluator.

You are an impartial evaluator.
Your task is to score a model's answer to a given question against a gold (reference) answer.

Scoring criteria:
- 0: Completely wrong or irrelevant
- 1: Partially correct but incomplete, vague, or missing key information
- 2: Completely correct and contextually accurate

Question: {question}
Gold Answer: {gold}
Model Answer: {prediction}

## Output Instructions
- Output only one line.
- The line must be in the exact format: "Score: X" where X is 0, 1, or 2.
- Do not generate explanations or additional text.

## Output Format
Score: 1
```

REALTALK and LoCoMo use benchmark-native prompt wrappers. The query-agnostic control changes only the reconstruction input field and reuses the validated reconstruction contract. The legacy Fact-style prompt above documents the historical factorial; it is not the clarified Generic v2 comparator in the main READ-stage analysis.

6.3 Additional Details for RQ1

6.3.1 Host Integration

RQ1 evaluates AffectLedger as an augmentation to five memory Hosts on the same CloneMem question populations. The reference retrieval Host supplies its BM25-selected traces directly to the downstream model. A-MEM, Mem0, Cognee, and MemoryBank retain their respective memory construction and retrieval procedures. AffectLedger uses the same supplementary interface across these Hosts.

MemoryBank constructs event-level memories and retrieves using MiniLM embeddings. Its evaluated configuration uses top-three retrieval and a 2048-token Host-context budget; the other Host-specific settings are given above rather than normalized to a common adapter.

6.3.2 Paired Cross-Host Statistics

Table 8 reports the paired results underlying the main cross-Host comparison. Rescue denotes a question answered correctly only by the augmented condition; harm denotes a question answered correctly only by Base.

Table 9: Complete CloneMem access and representation results (correct/N).

Condition	100k ($N = 186$)	500k ($N = 1001$)
Host only (Base)	160/186	845/1001
Host + Extra Retrieval	163/186	869/1001
Host + Selected Raw	165/186	901/1001
Host + Budget Raw	161/186	862/1001
Host + Generic v2	169/186	879/1001
Host + Relation-aware (Full)	171/186	876/1001

For MemoryBank-500k, the paired prediction vectors give 95 rescues and 33 harms. The exact two-sided McNemar test is $p = 3.8109453870922303 \times 10^{-8}$; the paired bootstrap interval (10,000 replicates, seed `20260918`) is $[+3.996, +8.392]$ percentage points.

6.4 Additional Results and Controls for RQ2

6.4.1 Complete Access and Representation Results

Table 9 reports exact correct-answer counts for the six conditions in the main READ-stage comparison. All conditions retain the same reference Host context and paired question populations. Extra Retrieval continues the Host's BM25 ranking. Selected Raw exposes AffectLedger's entire selected evidence sequence without reconstruction. Budget Raw packs complete selected observations in frozen order under a nominal added-context ceiling of 799 tokens (100k) or 926 tokens (500k), without splitting an observation; it may therefore expose a smaller evidence subset than Selected Raw. Generic v2 and Full reconstruct the same selected evidence, differing in their representation instructions. Generic v2 is the clarified-final version, not the earlier Fact-style control.

Generic v2 is evaluated on all 186/1001 aligned questions with the clarified-final prompt; the 30 clarified pilot reconstructions are incorporated into the final frozen lineage, while old-prompt rows are excluded. Selected evidence identity and order match canonical Full for all 1,187 questions. The 100k Generic outputs comprise 176 schema-valid outputs, 173 nonempty supplements and 13 empty supplements. At 500k, the corresponding counts are 990, 949, and 52. Empty or rejected supplements reuse Base; these are not transport failures. The verified final lineage reports no transport failures. Budget Raw yields a nonempty supplement in all questions; aggregate artifacts do not establish a zero retry count.

6.4.2 Historical Access Policy

AffectLedger retrieves eight semantic seeds from the cutoff-visible Ledger. Same-reference expansion uses resolved reference identities, while temporal expansion retrieves neighboring observations from the visible history. Both channels operate on historical observations not already included among the seeds. Candidates are combined and deduplicated. Selection preserves all seeds, attempts coverage from both expansion channels when available, and fills the remaining budget by query–evidence relevance. Duplicate source/span pairs and per-reference overrepresentation are constrained. The final selected sequence contains at most 16 observations.

Selected Raw, Generic v2, and Full are evaluated over the same selected evidence. Budget Raw is formed by packing complete records from that sequence under its token ceiling. Consequently, Budget Raw versus Full compares accuracy under similar downstream context sizes, but it does not isolate wording while holding the exposed evidence set fixed. Generic v2 versus Full holds the selected evidence fixed and compares representation policies.

6.4.3 Temporal Expansion and Reconstruction Factorial: Execution Audit

The historical factorial crossed temporal expansion (on/off) with legacy Fact-style or relation-aware reconstruction. It was not the clarified Generic v2 comparison in the main experiment. Table 10 retains the frozen 100k cells and distinguishes original from repaired 500k results. The latter are a repaired overlay on the archived execution, not a newly run complete experiment. Disabling

Table 10: Historical factorial (correct/N). Original 500k results are execution-confounded; the repaired overlay retains one unresolved Base fallback.

Scale	Lineage	Temporal	Fact-style	Relation-aware
100k	Frozen	On	165/186	171/186
	Frozen	Off	166/186	167/186
500k	Original	On	813/1001	876/1001
	Original	Off	885/1001	820/1001
500k	Repaired overlay	On	884/1001	876/1001
	Repaired overlay	Off	885/1001	893/1001

Table 11: Direct query-conditioning control. Values are accuracy (%).

Reconstruction	100k	500k
Query-agnostic	90.32	88.11
Full AffectLedger	91.94	87.51

temporal expansion sets its candidate budget to zero; the other configured pipeline stages remain unchanged.

The per-question interaction is $(R_{\text{on}} - F_{\text{on}}) - (R_{\text{off}} - F_{\text{off}})$, where R and F denote relation-aware and legacy Fact-style outcomes. The 100k estimate is +2.69 percentage points (paired 95% CI $[-3.23, +8.60]$). The original 500k point estimate of +12.79 points is not a valid positive-interaction finding: the original On–Fact-style cell contained 506 affected execution records, while the Off–Relation-aware cell contained 552 parse failures, mostly empty or whitespace provider responses. Their saved answers do not establish that the intended representation was actually supplied.

The repaired 500k overlay yields an interaction of -1.60 points (paired 95% CI $[-4.30, +1.10]$; 10,000 paired resamples, seed `20260918`). Each cell retains $N = 1001$ aligned scored outcomes, but one eligible Off–Relation-aware request remained unresolved after six attempts and uses a frozen Base fallback. Thus the repaired overlay resolves the bulk of the execution failures without constituting an unqualified all-repairs-complete factorial, and its CI crosses zero. No stable positive interaction is inferred from this audit.

6.4.4 Query-Agnostic Reconstruction

This control removes the current question from the reconstruction input while retaining question-conditioned evidence selection, reconstruction model, output contract, validation, and downstream answer procedure. Thus query-agnostic refers only to direct reconstruction conditioning; the evidence-access step remains question-conditioned. It tests the additional contribution of direct question conditioning after historical access.

The opposite numerical directions across scales do not establish a stable benefit from direct question conditioning at reconstruction, given the question-conditioned evidence selection already shared by both conditions.

6.4.5 Persistent Write-Time Consolidation and Evidence Revisitation

The Fixed baseline incrementally maintains a single persistent, query-agnostic longitudinal account using the same grounded local observations produced by canonical AffectLedger WRITE. Updates are organized by source unit, and the account available at a question cutoff is constructed only from observations visible by that cutoff. Fixed does not revisit earlier local observations at access time. Fixed+Raw supplies the persistent account and the same AffectLedger-selected raw evidence exposed by the Selected Raw condition. AffectLedger instead retains the local observations and constructs its longitudinal account from selected evidence at access time. All conditions retain the reference Host context and share the grounded observation substrate.

Table 12: Memory-lifecycle comparison. Accuracy is exact-choice percentage; added tokens are the mean supplementary downstream context.

Scale	Metric	Fixed	Fixed+Raw	Full
100k	Accuracy (%)	79.57	91.94	91.94
	Added tokens	8,766	10,317	799
500k	Accuracy (%)	85.51	89.21	87.51
	Added tokens	9,179	10,764	926

Table 13: Paired comparisons with Full AffectLedger; deltas are percentage points.

Scale	Comparator	Δ	95% CI
100k	Base	+5.91	[1.61, 10.75]
100k	Extra Retrieval	+4.30	[0.00, 9.14]
100k	Selected Raw	+3.23	[-1.61, 8.06]
100k	Budget Raw	+5.38	[+0.54, +10.75]
100k	Generic v2	+1.08	[-2.69, +4.84]
100k	Query-agnostic	+1.61	[-2.69, +5.91]
100k	Fixed	+12.37	[+6.45, +18.28]
500k	Base	+3.10	[1.00, 5.09]
500k	Extra Retrieval	+0.70	[-1.50, 2.90]
500k	Selected Raw	-2.50	[-4.50, -0.50]
500k	Budget Raw	+1.40	[-0.60, +3.50]
500k	Generic v2	-0.30	[-2.40, +1.80]
500k	Query-agnostic	-0.60	[-2.80, +1.50]
500k	Fixed	+2.00	[-0.30, +4.30]

At 100k, Fixed+Raw and Full each answer 171/186 questions correctly; the Fixed+Raw–Full paired comparison contains 11 Fixed+Raw-only and 11 Full-only correct answers (Fixed+Raw minus Full: 0.00 points, 95% CI $[-4.84, +4.84]$). At 500k, Fixed+Raw answers 893/1001 and Full 876/1001 correctly (67 Fixed+Raw-only and 50 Full-only; Fixed+Raw minus Full: +1.70 points, 95% CI $[-0.40, +3.90]$). For Fixed alone versus Full, Full minus Fixed is +12.37 points at 100k (95% CI $[+6.45, +18.28]$) and +2.00 points at 500k (95% CI $[-0.30, +4.30]$); the latter interval is expressed in the same direction as the point estimate.

These are comparisons of the evaluated lifecycle designs, not a timing-only ablation. Persistence, historical re-access, and reconstruction timing change together. Fixed+Raw is rendered in the final-answer supplement order *Fixed account, then selected raw evidence*, after the native Host context. The Fixed block includes cited supporting-evidence renderings; selected evidence is therefore represented again in the appended Raw block. This duplicate rendering contributes to the observed context size. No condition truncates its terminal answer context in the archived comparison. The design comparison does not isolate the contribution of repetition, nor does it measure end-to-end compute, provider billing, or latency.

6.4.6 Paired Mechanism Statistics

Table 13 presents the paired comparisons for the main READ-stage controls and other clean comparisons. Each delta is Full minus comparator in percentage points. Historical legacy Fact-style factorial results are documented separately above; they are not presented as the clarified Generic v2 control. Confidence intervals use 10,000 paired-question bootstrap replicates (seed `20260918`).

6.4.7 Answer-Level Correctness Transitions

Table 14 reports complete, aligned exact-choice outcomes, not a qualitative sample. In the Base–Selected Raw comparison, *target-only* means Selected Raw-only; in all other rows, target-only means Full-only. The categories are mutually exclusive and sum to N.

Table 14: Complete RQ2 correctness transitions. Columns give both correct, comparator-only, target-only, and neither correct.

Scale	Comparator → target	Both	Comp. only	Target only	Neither	N
100k	Base → Selected Raw	154	6	11	15	186
100k	Selected Raw → Full	158	7	13	8	186
100k	Budget Raw → Full	154	7	17	8	186
100k	Generic v2 → Full	163	6	8	9	186
500k	Base → Selected Raw	814	31	87	69	1001
500k	Selected Raw → Full	837	64	39	61	1001
500k	Budget Raw → Full	812	50	64	75	1001
500k	Generic v2 → Full	821	58	55	67	1001

Table 15: Mean added downstream-context tokens (complete paired populations).

Supplement	100k	500k
Base	0	0
Extra Retrieval	1,022.10	1,164.28
Selected Raw	1,549.13	1,584.00
Budget Raw	763.82	888.61
Generic v2	1,022.25	1,010.76
Relation-aware (Full)	798.80	925.74
Fixed account	8,766.48	9,179.44
Fixed+Raw combined	10,316.57	10,764.43

The transitions distinguish gains from selected evidence before reconstruction from subsequent representation effects. They do not independently establish why an individual answer changed; the trace-level inspection below provides descriptive, rather than causal, interpretation.

6.4.8 Context Cost

We count the actual rendered supplement appended to the unchanged Host's final-answer context with `tiktoken o200k_base`, excluding the native Host context, answer output, WRITE computation, and reconstruction-provider input/output. Table 15 reports complete-population mean added tokens. Actual realized tokens need not equal nominal ceilings because complete raw records are packed without splitting. In Fixed+Raw, the actual combined rendered string is counted, including both blocks and their separators; it is not reconstructed by adding rounded marginal means.

Selected Raw has median added tokens 1,554.5 (100k) and 1,582.0 (500k); Full has median 831.5 and 951.0. Compared with complete Selected Raw, Full reduces mean added context by 48.4% and 41.6% respectively. Compared with the evaluated Fixed+Raw combined supplement, it reduces added context by 92.3% and 91.4%, respectively. These are answer-time context measurements; they are not end-to-end latency, API usage, dollar costs, or a general estimate for all persistent-memory designs.

6.4.9 Qualitative Trace Audit

The first-pass, trace-grounded inspection contains 16 sampled transition headings corresponding to 15 distinct questions; one question occurs in two transition strata. Sampling uses aligned Host rendering, selected observations, reconstruction claims/citations, and downstream answers. The audit describes evidence recovered beyond Host context, explicit cross-time organization, omission or compression of selected details, changed emphasis, and unavailable or empty outputs. Its frequencies are descriptive of this inspection, not estimates over the complete CloneMem populations.

Provenance/identifier validation, claim-level semantic support, and answer-level correctness are separate notions. A validator PASS verifies that accepted claims satisfy the output and evidence-identifier contract; it does not independently verify that every proposition is semantically faithful. A supported reconstructed account can nevertheless omit or downweight a detail decisive for a downstream question, and the final answer can be incorrect despite a valid nonempty supplement. In

Table 16: Complete aggregate results for additional RQ3 benchmarks. Metrics are reported on their native scales.

Benchmark	Metric	N	Base	+AffectLedger	Δ
BEAM-128K	Native rubric	400	0.3579	0.3890	+0.0312
PersonaMem-128K	Accuracy	2727	0.6175	0.6036	-0.0139
REALTALK	Semantic judge	728	0.6470	0.6690	+0.0220
REALTALK	Lexical	728	0.2908	0.2894	-0.0014
ES-MemEval	Official F1	1427	0.3219	0.3182	-0.0037
ES-MemEval	Mean judge	1427	1.3966	1.3973	+0.0007
LoCoMo	Official aggregate	1986	0.5058	0.4580	-0.0478

Table 17: Complete native task-category accuracy on CloneMem. Δ denotes percentage-point change.

	CloneMem-100k				CloneMem-500k			
Category	N	Base	+AL	Δ	N	Base	+AL	Δ
Causal	28	71.43	92.86	+21.43	174	79.31	85.63	+6.32
Comparison	19	94.74	89.47	-5.26	160	89.38	93.75	+4.37
Counterfactual	37	86.49	86.49	0.00	151	72.19	80.13	+7.95
Inference	23	95.65	95.65	0.00	102	92.16	90.20	-1.96
Pattern	33	96.97	96.97	0.00	161	81.37	81.99	+0.62
Trajectory	33	72.73	90.91	+18.18	157	86.62	91.08	+4.46
Unanswerable	13	92.31	92.31	0.00	96	97.92	92.71	-5.21

the sample, Generic v2 failures include citations outside selected evidence and negative/uncertainty meta-claims; other errors involve empty output or incorrect answering after a valid nonempty supplement. Figure 3 illustrates three representative transformations without implying that these cases quantify the population-wide frequency of any mechanism.

6.5 Complete Results for RQ3

6.5.1 Benchmark-Level Results

Table 16 reports the complete benchmark-level comparisons, retaining the native metric for each evaluation.

6.5.2 CloneMem Capability Breakdown

Table 17 reports all native CloneMem task categories at both history scales.

6.5.3 BEAM Category Results

Each BEAM native category contains 40 examples.

6.5.4 PersonaMem Category Results

6.5.5 REALTALK Category Results

REALTALK uses the full available conversation history as its reference context. Table 20 reports the separate semantic-judge scores used for the capability analysis.

The semantic and lexical evaluations are distinct. The aggregate lexical score changes from 0.2908 to 0.2894; it is not interchangeable with the semantic-judge result.

Table 18: Complete category-level results on BEAM-128K. Values are the native rubric scores expressed on a 0–100 scale; Δ is the absolute point change.

Category	Base	+AffectLedger	Δ
Abstention	42.50	37.50	-5.00
Contradiction resolution	5.31	6.88	+1.56
Event ordering	17.84	17.43	-0.41
Information extraction	59.74	58.70	-1.04
Instruction following	46.25	47.50	+1.25
Knowledge update	26.25	41.25	+15.00
Multi-session reasoning	37.85	41.21	+3.36
Preference following	59.79	69.38	+9.58
Summarization	26.07	32.95	+6.89
Temporal reasoning	36.25	36.25	0.00

Table 19: Complete native question-type accuracy on PersonaMem-128K. Values are percentages; Δ denotes percentage-point change.

Question type	N	Base	+AffectLedger	Δ
Acknowledge latest user preferences	866	67.90	66.28	-1.62
Generalize to new scenarios	213	47.89	50.23	+2.35
Preference-aligned recommendations	349	53.58	49.57	-4.01
Recall user-shared facts	171	76.02	74.85	-1.17
Revisit reasons behind preference updates	269	86.62	86.25	-0.37
Suggest new ideas	518	35.52	32.24	-3.28
Track full preference evolution	341	76.25	77.71	+1.47

6.5.6 Statistical Summary

Table 21 summarizes the paired uncertainty estimates for the additional benchmark-level evaluations.

6.6 Qualitative Case Details

Figure 3 illustrates three distinct READ-stage properties: recovery of relevant historical evidence, organization of a fixed evidence set, and compact representation of selected history.

6.6.1 Selective-Access Rescue

The first example is drawn from CloneMem-500k (`0cd30bad-ebc1-47af-974d-cc5c04bf373a`), category counterfactual/opinion. The question asks whether the later three-year plan would have existed without the earlier realization that the problem was the self rather than the profession. The top-8 Host context contains the later operations commitment and three-year plan, but not the complete earlier career-choice-to-self transition. AffectLedger selects the following exact spans: *"Will I regret changing careers"* (2022-06-10), *"I'm starting to think that maybe the problem isn't actually my career choice."* (2022-07-02), *"I've realized the problem isn't the career; it's me."* (2022-07-11), and a later concern about persistence, support, and plan feasibility (2023-04-05). The Base answer is E and the augmented answer is B (the gold choice). This separates the Host context from the selected Ledger evidence and does not assert a source quote beyond the archived spans.

6.6.2 Relation-Aware Reconstruction Rescue

The second example comes from CloneMem-100k (`e539c6c9-c703-44e4-8e62-dd0d7b754617`). The question concerns a longer-term transition from exhaustion and breakdown to a more recent change in self-understanding and intention.

Table 20: REALTALK semantic-judge results by native category, expressed on a 0–100 scale.

Category	N	Base	+AL	Δ
Temporal reasoning	319	67.40	70.22	+2.82
Commonsense	108	54.63	57.41	+2.78
Multi-hop	301	65.45	66.78	+1.33

Table 21: Paired benchmark-level differences and bootstrap confidence intervals. Differences are +AffectLedger minus Base, expressed in the native metric.

Benchmark / metric	Difference	95% CI
BEAM native rubric	+0.0312	[+0.0026, +0.0589]
PersonaMem accuracy	-0.0139	[-0.0297, +0.0015]
REALTALK semantic	+0.0220	[-0.0082, +0.0522]
REALTALK lexical	-0.0014	[-0.0145, +0.0112]
ES-MemEval F1	-0.0037	[-0.0099, +0.0022]
ES-MemEval mean judge	+0.0007	[-0.0252, +0.0273]
LoCoMo aggregate	-0.0478	[-0.0610, -0.0352]

All three conditions use the same selected evidence. The Fact-style condition is the legacy factorial comparator, rather than Generic v2 in Table 2. Their downstream outcomes are shown in Table 22. Their downstream outcomes are shown in Table 22.

The relation-aware representation makes the phase structure explicit: earlier exhaustion and breakdown are followed by later understanding and intention. The example illustrates a representation effect over a fixed evidence set, rather than a difference in retrieval.

6.6.3 Compact Reconstruction

The third example is drawn from CloneMem-500k (`58f74ad8-cc77-4ebd-97dd-46c298aab146`). The question concerns how experiences of being understood changed the user's interpretation of communication and understanding within their family.

The selected evidence contains 16 observations across eight timestamps. Selected Raw supplies 1,553 added tokens, while relation-aware reconstruction produces a single claim supported by five observations across five timestamps, requiring 769 added tokens. The reconstruction validator is PASS, but its recorded semantic support status is `partial`; this is a contract-valid trace, not independent semantic-faithfulness verification.

The reconstruction reduces added context by 784 tokens (50.5%) while preserving the correct downstream answer. The reconstructed account links the experience of feeling understood to a later reconsideration of what understanding and communication could mean.

6.7 Computational Footprint and Reproducibility

6.7.1 Computational Footprint

We distinguish one-time Ledger construction, access-time reconstruction, and answer-time supplementary context. Table 24 reports observed SQLite storage and terminal WRITE outcomes. The *nonfailed outcomes* column includes units that produced no eligible affective observation or were skipped by provenance checks; it should not be interpreted as a count of successful model responses.

For canonical CloneMem WRITE, the persisted semantic-attempt records contain 1,609 attempts and 1,204 successful model responses over 1,048 source units at 100k, and 13,368 attempts and 12,591 successful model responses over 7,441 units at 500k. These attempt- and source-unit denominators are different. Additional deterministic processing and index operations are not converted into model-call estimates.

There is one successful terminal reconstruction per aligned question, *not* one attempt per question. The remaining 192 (100k) and 365 (500k) persisted attempts are additional/failed attempts. Their

Table 22: Representation comparison for the reconstruction-rescue case.

Representation	Answer
Selected Raw	Incorrect
Legacy Fact-style	Incorrect
Relation-aware	Correct

Table 23: Context reduction in the compact-reconstruction case.

Representation	Added tokens	Answer
Selected Raw	1553	Correct
Relation-aware	769	Correct

archived fields do not reliably separate transport retries from semantic retries. The final nonempty supplements number 153/186 and 928/1001, respectively. Canonical provider input/output token usage, reliable end-to-end latency, provider billing, and dollar cost are unavailable from the frozen fields. Consequently, we report only the directly observed attempt counts, SQLite footprint, and offline `o200k_base` added answer-time context statistics in Table 15; no end-to-end cost claim is made.

6.7.2 Run Lineage and Result Verification

The reported results correspond to frozen experiment configurations and identifiable terminal artifacts. The experiment records preserve benchmark splits, Host configurations, model roles, prompt versions, evaluation metrics, and per-question results where available.

The canonical CloneMem populations contain 186 and 1,001 questions at the two history scales. The access and representation comparisons reuse the same question populations and Host contexts; the factorial conditions preserve paired question alignment. Benchmark-specific native metrics are retained rather than converted into a common accuracy measure.

The supplementary evidence package includes anonymized aggregate results, release-safe paired score vectors where frozen fields were available, configuration and prompt snapshots, and a manifest connecting reportable results to their evaluated artifacts. Source code and full conversation corpora are not included. Unavailable fields are marked as such rather than reconstructed.

Prompt blocks distinguish recovered evaluated static instructions from request templates and unavailable wrappers. The clarified Generic v2 static version is identified separately from request-level hashes and from the historical Fact-style control.

Table 24: Persisted Ledger construction footprint. SQLite bytes exclude separately rebuildable indexes and caches.

Population	Source units	Nonfailed	Accepted units	Ledger entries	SQLite bytes
CloneMem-100k	1,048	1,046	848	4,177	3,899,392
CloneMem-500k	7,441	7,436	5,682	27,693	25,522,176
ES-MemEval	4,689	4,673	3,241	4,198	3,497,984

Table 25: Canonical READ reconstruction attempts and final supplements. Counts distinguish unique questions from persisted attempts.

Scale	Questions	Attempts	Successes	Empty supplements
100k	186	378	186	33
500k	1,001	1,366	1,001	73