GROUNDEDMEM: SOURCE-RESOLVED RETRIEVAL AND CONSERVATIVE REFUSAL FOR LONG-TERM CONVERSATIONAL MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term conversational memory requires recovering information from growing interaction histories. However, retrieval relevance does not imply evidential sufficiency: related memories may refer to the wrong participant, time, or event state, while missing support in a bounded retrieval set does not establish global unanswerability. Most long-term-memory systems do not explicitly report refusal quality, while those that do remain substantially weaker on this axis. We present GroundedMem, an auditable pipeline linking retrieval, source recovery, and selective answering. Compact Atomic memories serve as semantic indices; Atomic and direct-turn retrieval are fused and reranked; and selected candidates are resolved to complete, source-linked conversation turns. A controller refuses only when constrained mismatch checks fire without detected support; unresolved cases remain on the ordinary answer path. We develop the verifier prompt and hard decision policy on LongMemEval-S and transfer them to LoCoMo after normalizing candidate certificates, without LoCoMo-label tuning of the verifier. Relative to the published comparator rows included in our comparison, GroundedMem improves LoCoMo LLM-judge accuracy by 5.4 points and LongMemEval-S accuracy by 3.4 points. LoCoMo refusal F1 exceeds the reported REMem value by 9.3 points. In benchmark-internal paired comparisons, with refusal disabled and the Answer model, Judge protocol, prompt, and evidence budget fixed between arms, the complete evidence path exceeds Direct+EvoEmbedding by 2.97 points on LoCoMo and 1.60 points on LongMemEval-S. A separate paired LoCoMo ablation yields a net accuracy gain of 8.06 points from refusal, with newly incorrect answers accounting for 0.45% of answerable questions. Complete annotated source coverage nevertheless leaves substantial answer errors, motivating separate evaluation of discovery, evidence use, and answer decisions.

1 INTRODUCTION

Long-running interactions require language models to retain information from an expanding history and recover it when later requests depend on that information. Existing systems commonly extract or organize compact memories, retrieve items relevant to a query, and generate an answer from the retrieved context (Park et al., 2023; Chhikara et al., 2025). LoCoMo and LongMemEval have advanced this direction through cross-session question answering (Maharana et al., 2024; Wu et al., 2025). A central question within this process is whether a retrieved memory contains the evidence needed to answer the current question. This question connects the memory system's read interface to the reliability of its final response: successful access to a related item is only an intermediate outcome, not the answer decision itself.

Relevance differs from evidential support in two ways. First, a semantically related memory may refer to a different participant, time, or event state. For example, an application plan does not establish graduation, and experiences of different family members cannot be interchanged. Compact memories are useful retrieval cues but may omit context needed to interpret a fact. Second, when a bounded Top-K result lacks support, the fact may be absent from the history or may simply have been missed. Treating local absence as global unanswerability can turn a retrieval miss into a false refusal.

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

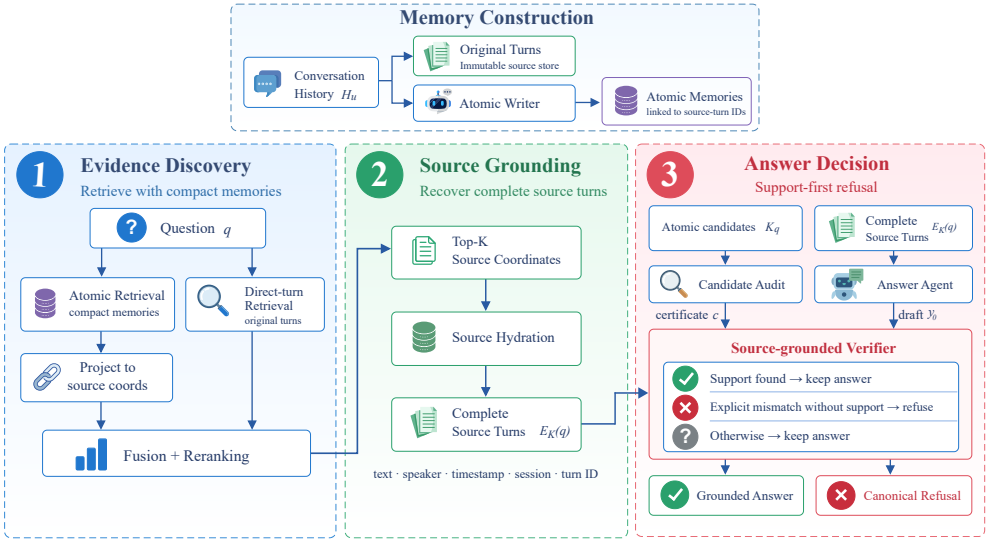


Figure 1: The GroundedMem architecture: dual-view discovery, complete source-turn grounding, and support-first answer verification.

These distinctions create a particular challenge for long-term memory. A compact fact can make a long conversation searchable, but the same abstraction may remove the speaker attribution, relative-time expression, or surrounding event description needed by a later question. Conversely, searching only the original turns preserves context but may miss facts expressed indirectly or across locally related statements. An effective system therefore needs both semantic access and a faithful source layer. It also needs a decision policy that does not collapse two different situations—failure to find support and discovery of a close but incompatible fact—into the same refusal signal.

We organize this problem around three questions. *Discovery:* which representations help recover the relevant history? *Grounding:* what information should actually reach the answer model? *Decision:* when does the available evidence justify constraining the response? The three questions have different observable outcomes. Source-turn coverage measures discovery; the serialized source context exposes grounding; and paired answer changes measure the effect of the decision layer. A single answer-accuracy score cannot reveal which stage accounts for a gain or an error.

GroundedMem organizes memory use around these distinctions. Atomic memories and original turns provide complementary retrieval views. After fusion and reranking, candidates are resolved through source coordinates to complete turns. A separate decision layer applies constrained support and mismatch checks to provenance-linked Atomic candidates. Ordinary absence remains unresolved and continues through the existing answer path. Retrieval representations discover information, source turns provide answer evidence, and the controller limits risks associated with specific mismatches.

Relative to the comparator rows in Table 1, GroundedMem improves LongMemEval-S answer accuracy by 3.4 points and LoCoMo LLM-judge accuracy by 5.4 points. Full source-turn recall is 91.91% at Top-10 on LongMemEval-S and 82.15% at Top-15 on LoCoMo. Under controller-off paired comparisons, the complete evidence path improves over Direct+EvoEmbedding by 2.97 points on LoCoMo and 1.60 points on LongMemEval-S. A separate full-cohort LoCoMo ablation attributes a net gain of 160 correct answers to the refusal module. We also report false-refusal and induced-error rates and analyze residual errors under complete annotated evidence coverage.

Our contributions are:

- an auditable memory pipeline that assigns distinct roles to semantic indexing, source recovery, and answer decisions;
- dual-view retrieval that resolves all routes to source-linked turns, followed by refusal that prioritizes detected support over restricted mismatch types; and
- controlled answer-side comparisons against Direct+EvoEmbedding on two benchmarks, together with coverage-conditioned analysis and a full-cohort paired ablation quantifying both the gains and harms of refusal.

2 Related Work

2.1 Long-term conversational memory

Long-term-memory systems differ in what they retain and how that information is recovered. Prior work tracks evolving user information across sessions (Bae et al., 2022); stores experiences with reflection, relevance, and recency (Park et al., 2023); and maintains persistent, updatable user memories (Zhong et al., 2024). Mem0 operationalizes scalable fact extraction, consolidation, updating, and retrieval, with a graph extension for relational recall (Chhikara et al., 2025). Later systems study granularity and adaptive retrieval (Pan et al., 2025; Tan et al., 2025), linked notes and graph organization (Gutiérrez et al., 2025; Xu et al., 2025), and hierarchical or structured retrieval (Banerjee et al., 2026; Yue et al., 2026). Controlled evidence also questions when graph structure is necessary (Hu et al., 2026b).

GroundedMem constructs an auditable evidence path: it combines Atomic and direct-turn retrieval, normalizes selected routes to complete source turns, and jointly evaluates source coverage, evidence use, and the answer-or-refuse decision. We use the public EvoEmbedding-4B model for contextual reranking (Nie et al., 2026).

2.2 Source grounding and evidence use

Attributed generation separates topical relevance from evidential support. AIS formalizes whether claims are attributable to identified sources (Rashkin et al., 2023); RARR retrieves attribution evidence and revises unsupported statements (Gao et al., 2023a); and ALCE jointly evaluates retrieval, answer correctness, and citation quality (Gao et al., 2023b). GroundedMem brings this distinction inside conversational memory. Provenance links materialize retrieved memories as complete turns with speaker and temporal context before generation. This enables separate evaluation of annotated source coverage and final answer quality.

2.3 Evidence sufficiency and selective answering

Retrieved context may be relevant; however, it may still be incomplete or incompatible with the question. FLARE triggers additional retrieval from low-confidence generations (Jiang et al., 2023), and Self-RAG learns when to retrieve and critique retrieved passages and generations (Asai et al., 2024). Sufficient Context separates inadequate context from failures to use adequate context and applies a learned sufficiency signal to abstention (Joren et al., 2025). LongMemEval includes abstention among its core abilities, while REMem evaluates refusal on unanswerable conversational questions (Wu et al., 2025; Shu et al., 2026). GroundedMem focuses on bounded memory retrieval: it restores answer context through provenance and forces refusal only for constrained mismatch evidence, while ordinary absence remains unresolved.

3 Method

Figure 1 summarizes the pipeline. Both benchmarks share retrieval, reranking, source restoration, and the Answer runtime; adapters provide available temporal metadata, Answer instructions, and refusal policies.

3.1 Memory Construction and Source Evidence

Let $H_u = (h_1, \ldots, h_N)$ be a user's timestamped turns and q a question. Model-based extraction produces Atomic memories $M_u = \{m_j\}_{j=1}^{L}$. Each item stores an ordered list of session and source-turn coordinates, defining $\pi(m_j)$. These links are retained during construction, so query-time restoration reads the recorded turns rather than rematching a summary by similarity. With Atomic and direct retrieval routes $R_M(q)$ and $R_H(q)$, the selected primary evidence is

$$E_K(q) = \mathrm{Select}_K(\mathrm{Prior}_M(\mathrm{Rerank}(\mathrm{Fuse}(\mathrm{Resolve}(R_M(q), \pi), R_H(q))))) \,. \tag{1}$$

Each selected turn retains its complete text, speaker, timestamp, session, coordinate, and retrieval routes. A shared formatter constructs the answer payload, including evidence-derived task instructions and question time when supplied by the benchmark. Atomic text supports retrieval and controller checks; complete source turns provide primary answer evidence. Appendix A specifies memory writing and the full inference algorithm.

3.2 Dual-View Evidence Discovery

Atomic retrieval searches localized propositions; direct retrieval preserves wording and context that extraction may omit. The Atomic route expands its top 30 memories to source turns in memory-rank order and deduplicates each coordinate at first occurrence. The direct route first fuses BM25 and dense rankings of original turns by equal-weight RRF with constant 60. The Atomic and direct routes then contribute source-turn rankings through equal-weight RRF:

$$s_{\mathrm{fuse}}(h) = \frac{\mathbf{1}[h \in R_M^H]}{60 + r_M(h)} + \frac{\mathbf{1}[h \in R_H]}{60 + r_H(h)}, \tag{2}$$

Here $R_M^H = \mathrm{Resolve}(R_M, \pi)$, and r_M, r_H, and r_E denote one-based ranks in their respective routes. An absent turn has rank $+\infty$ and contributes zero. EvoEmbedding reranks the top 100 fused candidates. With reranked position $r_E(h)$, the final score retains an Atomic prior:

$$s_{\mathrm{final}}(h) = \frac{\mathbf{1}[h \in R_E]}{10 + r_E(h)} + \frac{0.5\,\mathbf{1}[h \in R_M^H]}{10 + r_M(h)}. \tag{3}$$

This prior does not require an item to occur in both views. The top K complete turns form the primary evidence, with $K = 10$ on LongMemEval-S and $K = 15$ on LoCoMo. Reference answers, answerability labels, and gold sources are excluded from ranking.

3.3 Selective Answering with Support Precedence

Let G_b denote the benchmark-specific candidate adapter and τ_b its task contract. The adapter proposes a certificate c_b from a bounded Atomic set $B_M(q)$, while the Answer Agent produces a draft y_0 from the selected complete source turns E_K. For proposed cases, the shared verifier returns $v_b = V(q, \tau_b, c_b, y_0, E_K)$. Write ν_b, κ_b, and S_b for its verdict, conflict type, and support IDs. The final decision is

$$d_b(q, c_b, y_0, E_K) = \begin{cases} \textsc{Refuse}, & c_b \neq \varnothing \ \wedge \ \nu_b = \texttt{VERIFIED_CONFLICT} \\ & \wedge\, \kappa_b = \texttt{TARGET_FRAME_SUBSTITUTION} \\ & \wedge\, S_b = \varnothing \\ & \wedge\ \mathrm{op}(\tau_b) \notin \mathcal{A}, \\ \textsc{KeepDraft}, & \text{otherwise.} \end{cases} \tag{4}$$

Positive support vetoes refusal. Unresolved cases keep the draft, which may itself abstain; only a verified target-frame substitution without support can force refusal outside advice, preference, and recommendation operations. LoCoMo's LLM-based fallback marker classifies three Atomic candidates. The candidate adapters differ by benchmark, while certificate normalization lets them share the source-grounded verifier and decision rule. Prompt contracts appear in Appendix C.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and metrics. LongMemEval-S has 500 questions, including 470 answerable and 30 unanswerable (Wu et al., 2025). LoCoMo has 1,986 questions over ten conversations; the REMem-compatible policy treats 1,540 non-Category-5 questions as answerable and 446 Category-5 questions as unanswerable (Maharana et al., 2024; Shu et al., 2026). We report LoCoMo token F1, BLEU-1, and LLM-judge accuracy, and LongMemEval-S overall and category accuracy. Full source-turn recall@K requires every annotated source turn to appear among the K selected turns; its cohorts comprise 470 and 1,535 answerable questions, respectively. Five LoCoMo questions lack valid source annotations.

Refusal precision, recall, and F1 use benchmark answerability labels. Controller and final false-refusal rates count answerable questions refused by the gate or the final system, respectively; induced-error rate counts paired correct-to-wrong changes. All three use the total answerable cohort as denominator.

Controlled comparisons. Retrieval comparisons progress from individual routes to Direct+EvoEmbedding, fusion, and the Atomic prior. The paired answer comparison disables forced refusal and contrasts Direct+EvoEmbedding with Atomic+Direct+EvoEmbedding+Prior. Within each benchmark, both arms fix data, Answer model and prompt, evidence budget, and Judge protocol. A separate LoCoMo ablation compares no refusal with verifier-controlled selection of previously generated controller-on and controller-off outputs. We report paired changes, exact two-sided McNemar tests, and 20,000-sample history-cluster bootstrap intervals conditional on the observed model outputs.

Implementation and evaluation. BM25 and Qwen3-Embedding-0.6B retrieve candidates; EvoEmbedding-4B reranks them (Nie et al., 2026). Answer generation and judging use GPT-4.1-mini; semantic candidate checks and the shared verifier use DeepSeek-v4-flash. LoCoMo follows REMem's metrics and Judge prompt, while LongMemEval-S follows its category-specific Judge rules. Reference answers, benchmark categories, answerability labels, and gold sources are excluded from runtime inputs. Controller query types are derived from the question. Appendix D records data versions, configuration, and label isolation; Appendix C.5 reports a two-annotator human audit of both Judges.

4.2 MAIN RESULTS

Table 1 compares overall, category, and refusal quality. RAG denotes retrieval-augmented generation; REMem-I and REMem-S are the named variants of the cited system. GroundedMem achieves 43.9 F1, 34.7 BLEU-1, and 82.9% LLM-J on LoCoMo, with refusal F1 of 73.3%. LongMemEval-S accuracy is 85.4% (427/500), including 86.4% on answerable and 70.0% on unanswerable questions. Refusal detection and answer correctness are scored separately.

4.3 RETRIEVAL AND EVIDENCE-PATH COMPARISONS

Source discovery. Table 2 shows that LoCoMo dual-view fusion raises full recall from Direct+EvoEmbedding's 76.74% to 81.43%, and the Atomic prior to 82.15%. On LongMemEval-S, fusion and Direct+EvoEmbedding both reach 90.85%; the prior raises recall to 91.91%. Gains therefore differ across benchmarks. Appendix B.1 reports complete budget curves.

Answer-side comparison. With refusal disabled, the complete path improves LoCoMo accuracy from 72.16% to 75.13%, with a history-bootstrap interval of $[+1.21, +4.72]$ points (Table 3). LongMemEval-S changes from 82.80% to 84.40%, with an interval of $[-1.00, +4.20]$ and $p = 0.302$; this does not establish a reliable answer-accuracy gain. The comparison changes Atomic retrieval and its prior together and therefore measures the complete path, not either component's marginal effect. Appendix B.2 records the breakdowns and evaluation boundary.

Table 1: Answer-quality comparison (%). In (a), F1 is token F1, BLEU-1 is unigram BLEU, and LLM-J is LLM-judge accuracy; in (c), precision, recall, and F1 treat refusal as the positive class, and Refusals is a count. SH/MH/T/OD/ADV denote LoCoMo single-hop, multi-hop, temporal, open-domain, and adversarial LLM-judge accuracy. LongMemEval-S categories are SSU (single-session user), MS (multi-session), SSP (single-session preference), TR (temporal reasoning), KU (knowledge update), and SSA (single-session assistant). All percentages are rounded to one decimal; bold marks column maxima.

(a) LoCoMo answer quality and category LLM-J

Group	Method	SH	MH	T	OD	ADV	F1	BLEU-1	LLM-J
Reference	Full Context	84.4	75.2	54.2	88.3	55.0	37.8	28.6	76.7
Embedding	Qwen3-Embedding-8B (Zhang et al., 2025)	64.2	57.1	45.8	67.3	66.8	35.3	28.9	64.2
Embedding	NV-Embed-v2 (Lee et al., 2025)	77.9	64.9	51.0	82.6	61.2	39.6	31.0	73.0
Memory	Mem0 (Chhikara et al., 2025)	32.7	52.1	46.9	57.2	46.6	25.1	18.0	49.7
Memory	Graphiti (Rasmussen et al., 2025)	30.8	53.6	45.8	45.0	**83.0**	33.7	28.9	52.5
Memory	HippoRAG 2 (Gutiérrez et al., 2025)	72.0	72.3	56.3	81.7	65.7	39.0	30.8	74.0
Memory	REMem-I (Shu et al., 2026)	81.3	70.2	56.3	83.5	66.8	42.4	32.7	76.2
Memory	REMem-S (Shu et al., 2026)	86.0	69.5	**63.5**	85.1	65.3	41.3	31.5	77.5
Memory	GroundedMem	**89.4**	**83.3**	61.5	**90.7**	67.7	**43.9**	**34.7**	**82.9**

(b) LongMemEval-S accuracy with GPT-4.1-mini

Method	SSU	MS	SSP	TR	KU	SSA	Overall
Full Context	85.7	51.1	16.7	60.2	76.9	**98.2**	65.6
RAG	82.9	54.9	86.7	67.7	80.8	94.6	72.6
PREMem (Kim et al., 2025)	92.9	57.1	36.7	59.4	84.6	12.5	60.8
Mem0 (Chhikara et al., 2025)	94.3	66.9	86.7	75.9	87.2	96.4	80.8
Nemori (Nan et al., 2025)	90.0	55.6	86.7	72.2	79.5	92.9	74.6
EverMemOS (Hu et al., 2026a)	**100.0**	**78.5**	**96.7**	71.2	87.2	78.6	82.0
GroundedMem	**100.0**	75.9	53.3	**85.0**	**93.6**	96.4	**85.4**

(c) LoCoMo refusal quality over all 1,986 questions

Method	Refusals	Precision	Recall	F1
Graphiti (Rasmussen et al., 2025)	954	38.9	**83.6**	53.1
Mem0 (Chhikara et al., 2025)	90	40.0	8.1	13.5
REMem (Shu et al., 2026)	344	73.3	56.8	64.0
GroundedMem	323	**87.3**	63.2	**73.3**

Table 2: Full source-turn recall (%) under fixed primary-turn budgets. @K means all annotated source turns occur among the top K primary turns; the evaluated answerable cohorts are LoCoMo $N = 1{,}535$ and LongMemEval-S $N = 470$.

Retrieval route	LoCoMo@15	LongMemEval@10
Atomic	76.42	88.30
Direct Turn	60.52	77.87
Direct + EvoEmbedding	76.74	90.85
Atomic + Direct + Evo	81.43	90.85
Atomic + Direct + Evo + Prior	**82.15**	**91.91**

Efficiency boundary. Following REMem, we separate indexing from inference and report time/memory and token/cost in two matched views. Indexing requires 580 and 51,245 seconds on LoCoMo and LongMemEval-S. The measured local Direct+EvoEmbedding fusion throughput is 5.76 and 14.13 seconds per question, with peak host memory of 17.81 and 20.98 GiB, respectively. The frozen indexing-plus-inference traces account for 26.65M and 283.89M tokens. Appendix D.3 defines each boundary and reproduces the REMem-style tables.

Coverage-conditioned diagnosis. Despite complete annotated-source coverage, 46 of 432 LongMemEval-S answers and 109 of 1,261 LoCoMo answers remain incorrect. Full coverage also occurs in 6 of 8 and 26 of 41 final false refusals, respectively. These errors require examining evidence use after retrieval; Appendix B.3 provides the complete contingency tables.

4.4 Refusal Intervention

On LoCoMo, verifier-controlled refusal raises correct answers from 1,486 to 1,646 (+8.06 points; Table 4), repairing 167 and harming seven relative to the no-refusal pair. Unanswerable accuracy

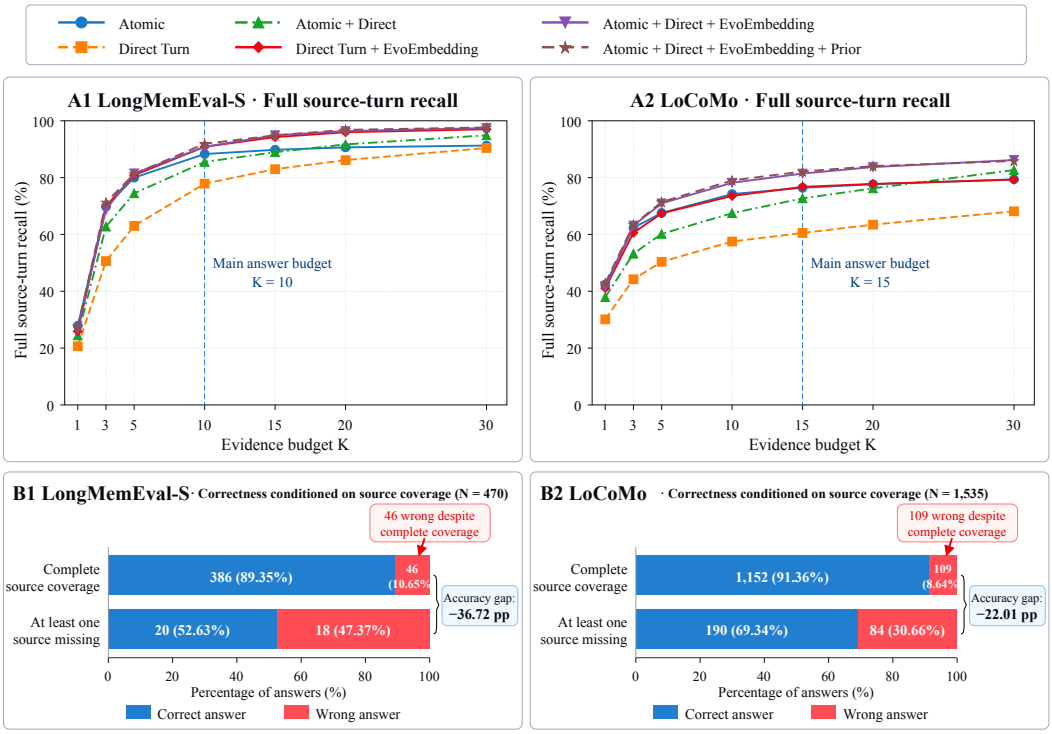


Figure 2: Retrieval coverage and answer correctness. Top: complete annotated source-turn recall at $K \in \{1, 3, 5, 10, 15, 20, 30\}$ over 470 answerable LongMemEval-S questions and 1,535 LoCoMo answerable questions with valid source annotations; dashed vertical lines mark the main answer budgets ($K = 10$ for LongMemEval-S and $K = 15$ for LoCoMo). Bottom: verifier-selected answer outcomes conditioned on whether all annotated source turns are present. Full coverage is diagnostic but not sufficient: 155 answers remain wrong despite complete coverage.

Table 3: Controller-off paired answer results. Both arms use the same data, Answer and Judge models, prompts, and evidence budget. Δ pp is the accuracy difference in percentage points; W→C and C→W count paired wrong-to-correct and correct-to-wrong changes for the complete path relative to Direct+EvoEmbedding. p is the two-sided exact McNemar value.

Benchmark	Direct+Evo	Complete path	Δ pp	W→C / C→W	p
LoCoMo	1,433/1,986	**1,492/1,986**	+2.97	136 / 77	6.40×10^{-5}
LongMemEval-S	414/500	**422/500**	+1.60	27 / 19	0.302

rises from 30.27% to 67.71%, while answerable accuracy changes from 87.73% to 87.27%. The verifier forcibly refuses 12 answerable questions (0.78%); seven become newly incorrect (0.45%).

On LongMemEval-S, verifier-controlled refusal changes 422/500 to 427/500, repairing five answers without newly incorrect cases ($p = 0.0625$). It forcibly refuses five unanswerable questions and no answerable questions. Appendix B.4 reports controller and final-output counts separately.

5 Discussion

5.1 Discovery: Complementary Retrieval Views

Atomic memories index semantic cues for paraphrased or distributed events; Direct Turn retrieval preserves original wording, rare entities, and speaker context. Adding Atomic candidates and its prior to Direct+EvoEmbedding raises full source-turn recall from 76.74% to 82.15% on LoCoMo ($K = 15$) and from 90.85% to 91.91% on LongMemEval-S ($K = 10$). Without the prior,

Table 4: Paired LoCoMo refusal ablation. Final false-refusal counts are 30/1,540 with refusal off and 41/1,540 with the verifier; rates include Answer Agent abstentions. Acc., Ans., and Unans. are overall, answerable, and unanswerable accuracy; False-ref. is the final false-refusal rate over 1,540 answerable questions (all rates in %). Lower false-refusal rates are better.

Refusal	Correct	Acc.	Ans.	Unans.	False-ref.
Off	1,486/1,986	74.82	**87.73**	30.27	**1.95**
Verifier	**1,646/1,986**	**82.88**	87.27	**67.71**	2.66

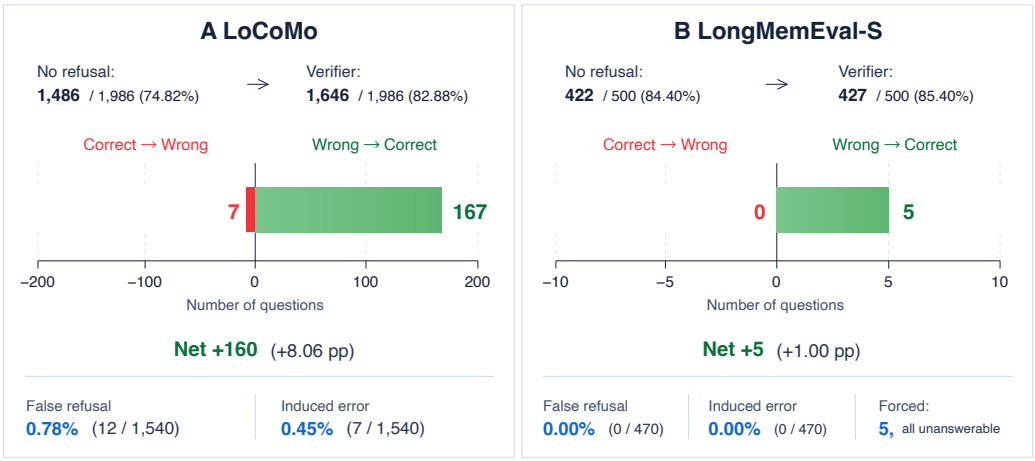


Figure 3: Paired effect of verifier-controlled refusal. Red leftward bars show correct-to-wrong changes and green rightward bars show wrong-to-correct repairs (LoCoMo: 7 versus 167; LongMemEval-S: 0 versus 5). Counts above compare correct answers with refusal off and on; the lower false-refusal and induced-error rates use answerable-question denominators.

LongMemEval-S remains at 90.85%: a larger candidate pool does not ensure that complementary evidence survives Top-K selection. The prior controls ranking, not evidence.

5.2 Grounding: Complete Source-Turn Restoration

Compact memories can surface a buried fact but also omit a relevant detail. Figure 4 contrasts two LongMemEval-S questions: an Atomic facet exposes a Thrive Market purchase embedded in meal-planning talk, whereas a Direct Turn preserves the Europe-trip time cue omitted by its Atomic summary. Fusion and reranking recover the complete source-turn sets (4/4 and 2/2), which GroundedMem restores before answering. Top-10 coverage is annotated post hoc, not provided to the retrievers.

With refusal disabled and Answer/Judge settings matched, the complete path improves over Direct+EvoEmbedding by 2.97 points on LoCoMo and 1.60 points on LongMemEval-S; the latter is not statistically established. Atomic retrieval and its prior change jointly, so this comparison does not isolate their answer-level effects. Complete source coverage also does not guarantee correct evidence use (Figure 2).

5.3 Decision: Support-First Refusal

GroundedMem considers refusal only after a source-grounded draft. Positive support vetoes forced refusal; unresolved conflicts fail open; preference and recommendation operations are excluded. Thus, missing local evidence is not treated as global unanswerability. The verifier repairs 167 LoCoMo answers while harming seven, and repairs five LongMemEval-S answers without new errors (Figure 3). Refusal acts as a constrained correction, not a default response to incomplete retrieval.

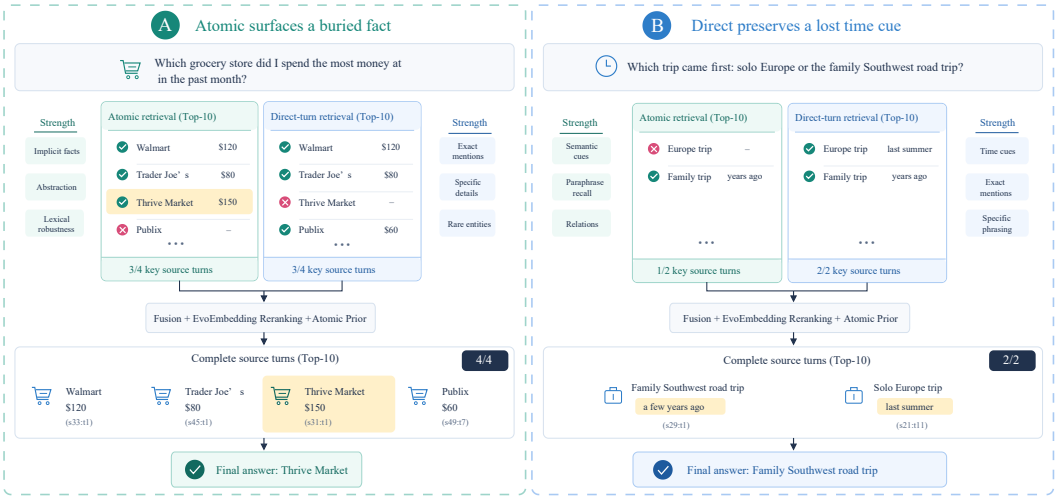


Figure 4: Two LongMemEval-S examples of complementary retrieval strengths. Atomic makes a purchase fact in a multi-topic turn searchable (left); Direct Turn retains a temporal cue omitted by the corresponding Atomic memory (right). Source-turn Top-10 coverage is annotated post hoc.

6 Limitations of the Memory System

The Atomic index is lossy, and source restoration cannot recover turns missed by both routes. Fixed budgets cannot adapt to question type; the offline store lacks provenance-aware updates and deletion; and source turns do not distinguish user preferences from assistant suggestions, which may contribute to preference errors. Missing support in Top-K does not prove unanswerability, so refusal fails open outside verified conflicts. Our refusal results reuse frozen Answer/Judge outputs on these cohorts; unseen-history and fresh end-to-end performance remain untested. Adaptive retrieval and auditable memory updates remain future work.

7 Conclusion

GroundedMem connects Atomic semantic discovery, complete source-turn restoration, and support-first refusal. Paired comparisons and coverage-conditioned diagnosis separate evidence access from answer quality and quantify refusal's benefits and costs. Future work will examine source-preserving updates and additional retrieval for unresolved questions.

AI Use Statement. AI tools were used to search for and discover related work.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25f7be9694d7b32d5cc670927b8091e1-Abstract-Conference.html.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. Keep me updated! memory management in long-term conversations. In *Findings of EMNLP*, pp. 3769–3787, 2022. doi: 10.18653/v1/2022.findings-emnlp.276.

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 *Annual Meeting of the Association for Computational Linguistics*, pp. 16470–16489, 2026. doi: 10.18653/v1/2026.acl-long.749.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *European Conference on Artificial Intelligence*, pp. 2993–3000, 2025. doi: 10.3233/FAIA251160.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of ACL*, pp. 16477–16508, 2023a. doi: 10.18653/v1/2023.acl-long.910.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of EMNLP*, pp. 6465–6488, 2023b. doi: 10.18653/v1/2023.emnlp-main.398.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515, 2025.

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. *arXiv preprint arXiv:2601.02163*, 2026a. URL https://arxiv.org/abs/2601.02163.

Sen Hu, Yuxiang Wei, Jiaxin Ran, Xueran Han, Zhiyuan Yao, Huacan Wang, Ronghao Chen, and Lei Zou. Does memory need graphs? a unified framework and empirical analysis for long-term dialog memory. In *Proceedings of ACL*, pp. 26758–26782, 2026b. doi: 10.18653/v1/2026.acl-long.1232.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of EMNLP*, pp. 7969–7992, 2023. doi: 10.18653/v1/2023.emnlp-main.495.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html.

Sangyeop Kim, Yohan Lee, Sanghwa Kim, Hyunjong Kim, and Sungzoon Cho. Pre-storage reasoning for episodic memory: Shifting inference burden to memory for personalized dialogue. *arXiv preprint arXiv:2509.10852*, 2025. URL https://arxiv.org/abs/2509.10852.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lgsyLSsDRe.

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 ACL*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747.

Jiayan Nan, Wenquan Ma, Wenlong Wu, and Yize Chen. Nemori: Self-organizing agent memory inspired by cognitive science. *arXiv preprint arXiv:2508.03341*, 2025. URL https://arxiv.org/abs/2508.03341v3.

Chang Nie, Chaoyou Fu, Junlan Feng, and Caifeng Shan. EvoEmbedding: Evolvable representations for long-context retrieval and agentic memory. *arXiv preprint arXiv:2606.21649*, 2026. URL https://arxiv.org/abs/2606.21649.

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 *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e56f394bbd4f0ec81393d767caa5a31b-Abstract-Conference.html.

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 *ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486.

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.

Yiheng Shu, Padmaja Jonnalagedda, Xiang Gao, Bernal Jimenez Gutierrez, Weijian Qi, Kamalika Das, Huan Sun, and Yu Su. REMem: Reasoning with episodic memory in language agent. In *International Conference on Learning Representations*, pp. 55958–55987, 2026. URL https://openreview.net/forum?id=fugnQxbvMm.

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 ACL*, pp. 8416–8439, 2025. doi: 10.18653/v1/2025.acl-long.413.

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

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. 17577–17604, 2025. doi: 10.52202/085713-0593.

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 ACL*, pp. 35237–35254, 2026. doi: 10.18653/v1/2026.acl-long.1627.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

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

APPENDIX OVERVIEW

Following the main paper, the appendix is organized by function: formal process, detailed results, prompt contracts, and implementation details.

Appendix A
Memory objects, provenance, retrieval indices, and algorithms.

Appendix B
Detailed retrieval, answer-path, category, coverage, and refusal results.

Appendix C
Atomic-writing, Answer, refusal, and Judge interfaces.

Appendix D
System configuration, label isolation, evaluation cohorts, metrics, and runtime.

A FORMAL DEFINITION

GroundedMem maintains two complementary memory views: original conversation turns and compact Atomic memories bound to those turns. The stored system consists of the history H_u, Atomic store M_u, provenance map P, and retrieval indices $\mathcal{I}_u$. Given a question q, it returns a source-grounded answer or canonical refusal $\perp$.

Memory objects. For user u, the ordered history is $H_u = (h_1, \ldots, h_N)$, with source turns $h_i = (\mathrm{id}_i, \mathrm{session}_i, s_i, \tau_i, x_i)$. Here id_i and $\mathrm{session}_i$ identify the original turn and session; s_i, τ_i, and x_i are its speaker, timestamp, and complete text. Define its source coordinate as $c_i = (\mathrm{session}_i, \mathrm{id}_i)$. The Atomic view is $M_u = \{m_j = (z_j, P_j, \mu_j)\}_{j=1}^{L}$, where z_j is a compact, independently retrievable facet, P_j is its nonempty ordered list of source coordinates in the saved Atomic payload, and μ_j contains temporal, entity, and retrieval metadata. Atomic text provides a semantic index; the linked original turns provide the primary answer evidence.

Provenance relation. The map $P : m_j \mapsto P_j$ records which source turns support each memory. Equivalently, its relation is $\mathcal{E}_P = \{(m_j, h_i) : c_i \in P_j\}$. Source adapters bind these identifiers to the original history before evaluation, and the mapping remains fixed during retrieval. Resolving a selected coordinate restores its complete source record, $\mathrm{Restore}(h_i) = (\mathrm{id}_i, \mathrm{session}_i, s_i, \tau_i, x_i, \mathrm{route}_i)$, including its retrieval routes. This operation follows the stored provenance binding and preserves speaker and temporal context.

Retrieval indices. We maintain lexical and dense indices over both Atomic memories and raw turns. Atomic retrieval uses a question-derived frame and optional subqueries, combines lexical and dense candidates, and contextually reranks them to obtain $R_M(q)$. Its first 30 memories are projected to source coordinates as $R_M^H(q) = \mathrm{Resolve}\Big(\mathrm{FirstUnique}\Big(\bigoplus_{m \in R_M(q)[:30]} P(m)\Big), H_u\Big)\,[:\,100]$, where concatenation preserves Atomic rank and each $P(m)$'s saved order; FIRSTUNIQUE keeps the first occurrence of each coordinate. Thus the Top-100 cutoff follows the frozen payload order, not a timestamp sort. Let R_B and R_D be the Top-100 BM25 and Qwen3-Embedding-0.6B rankings over raw turns. For any ranking R, let $r_R(h)$ be its one-based rank when $h \in R$ and $+\infty$ otherwise. Hence an absent route contributes zero under $1/(c + \infty) = 0$. Direct retrieval, cross-view fusion, and the final Atomic prior use

$$
\begin{aligned}
s_{\mathrm{direct}}(h \mid q) &= \frac{\mathbf{1}[h \in R_B]}{60 + r_{R_B}(h)} + \frac{\mathbf{1}[h \in R_D]}{60 + r_{R_D}(h)}, \\
s_{\mathrm{fuse}}(h \mid q) &= \frac{\mathbf{1}[h \in R_M^H]}{60 + r_M(h)} + \frac{\mathbf{1}[h \in R_H]}{60 + r_H(h)}, \\
s_{\mathrm{final}}(h \mid q) &= \frac{\mathbf{1}[h \in R_E]}{10 + r_E(h)} + 0.5\,\frac{\mathbf{1}[h \in R_M^H]}{10 + r_M(h)}.
\end{aligned}
\tag{5}
$$

R_H retains the first 100 turns under s_{direct}; $r_M = r_{R_M^H}$, $r_H = r_{R_H}$, and $r_E = r_{R_E}$. EvoEmbedding-4B reranks the first 100 under s_{fuse} to obtain R_E. Ranking $R_E \cup R_M^H$ by s_{final} yields R_P. The first K turns are restored as E_K, with $K = 10$ on LongMemEval-S and $K = 15$ on LoCoMo.

Algorithms. Algorithms S1 and S2 specify memory indexing and query-time answering. A benchmark adapter first proposes a refusal candidate, the Answer Agent produces a draft from restored turns, and the shared source-grounded verifier checks the candidate against that draft and the restored turns. Equation 4 maps the verifier output to either canonical abstention or retention of the draft. The Answer formatter uses the benchmark instructions and adds question time only when supplied by the benchmark. Reference answers and evaluation labels are excluded from both algorithms. Writing settings and statistics appear in Appendix D; prompt contracts appear in Appendix C.

Algorithm S1 Memory indexing

Require: Raw history H_u; fixed Atomic-writing contract W
Ensure: Atomic store M_u, provenance map P, retrieval indices $\mathcal{I}_u$

$M_u \leftarrow \emptyset$; $P \leftarrow \emptyset$
for each session-bounded writing batch B in H_u **do**
  $F \leftarrow \textsc{AtomicWrite}(B; W)$
  for each facet $m = (z, P_m, \mu)$ in F **do**
    Bind P_m to original turn and session coordinates
    Store m in M_u with $P(m) \leftarrow P_m$
  end for
end for
Freeze the source-coordinate map P
$\mathcal{I}_u \leftarrow \textsc{BuildLexicalDenseIndices}(M_u, H_u)$
return $(M_u, P, \mathcal{I}_u)$

Algorithm S2 Grounded answering

Require: Question q; history H_u; store M_u; map P; indices $\mathcal{I}_u$; benchmark b
Ensure: Answer $\hat{y} \in \mathcal{Y} \cup \{\perp\}$

$R_M \leftarrow \textsc{AtomicRetrieve}(q, M_u, \mathcal{I}_u)[:30]$
$R_M^H \leftarrow \textsc{Resolve}(\textsc{FirstUnique}(\bigoplus_{m \in R_M} P(m)), H_u)[:100]$
$(R_B, R_D) \leftarrow \textsc{BM25DenseRetrieve}(q, H_u, \mathcal{I}_u; 100)$
$R_H \leftarrow \big(\text{rank}_h\, s_{\text{direct}}(h \mid q)\big)[:100]$
$R_F \leftarrow \big(\text{rank}_h\, s_{\text{fuse}}(h \mid q)\big)[:100]$
$R_E \leftarrow \textsc{EvoRerank}(q, R_F)$
$R_P \leftarrow \text{rank}_{h \in R_E \cup R_M^H}\, s_{\text{final}}(h \mid q)$
$K \leftarrow 10$ if $b =$ LongMemEval-S, otherwise 15
$E_K \leftarrow \textsc{RestoreSourceTurns}(R_P[:K], H_u)$
$c \leftarrow \textsc{CandidatePolicy}_b(q, R_M)$
$y_0 \leftarrow \textsc{Answer}(q, E_K; b)$
$d \leftarrow \textsc{SourceGroundedVerify}(q, c, y_0, E_K)$
return $\perp$ if $d = \textsc{Refuse}$; otherwise y_0

B Detailed Results

B.1 Complete retrieval curves

Table 5: LongMemEval-S full source-turn recall (%) over 470 answerable questions. @K means all annotated source turns appear among the top K primary turns; Evo abbreviates EvoEmbedding, and Full + Prior denotes Atomic + Direct + EvoEmbedding + Atomic prior.

Route	@1	@3	@5	@10	@15	@20	@30
Atomic	**27.87**	69.79	80.00	88.30	89.79	90.64	91.28
Direct	20.64	50.64	62.98	77.87	82.98	86.17	90.43
Atomic + Direct	24.47	62.77	74.47	85.53	88.94	91.70	94.89
Direct + Evo	25.96	70.00	80.85	90.85	94.26	95.96	97.02
Atomic + Direct + Evo	26.60	68.72	**81.49**	90.85	**94.89**	96.38	97.45
Full + Prior	27.45	**71.28**	81.28	**91.91**	**94.89**	**96.81**	**97.66**

Table 6: LoCoMo full source-turn recall (%) over 1,535 answerable questions with valid source annotations. @K, Evo, and Full + Prior follow the definitions in the preceding table.

Route	@1	@3	@5	@10	@15	@20	@30
Atomic	42.35	62.35	67.56	74.20	76.42	77.65	79.41
Direct	30.10	44.23	50.36	57.52	60.52	63.45	68.14
Atomic + Direct	37.92	53.22	60.20	67.49	72.70	76.22	82.67
Direct + Evo	41.11	60.59	67.43	73.55	76.74	77.79	79.28
Atomic + Direct + Evo	41.56	63.13	70.94	78.18	81.43	83.71	**86.19**
Full + Prior	**42.87**	**63.45**	**71.40**	**79.02**	**82.15**	**84.04**	85.86

The curves measure source-turn recall across candidate budgets; the primary answer budgets are @10 and @15. The LoCoMo cohort includes only questions with valid source-turn annotations.

B.2 Controller-off paired answer comparison

The answer-side comparison uses controller-off outputs in both arms. The Direct+EvoEmbedding arm removes the Atomic retrieval route and Atomic prior; the complete arm retains Atomic+Direct fusion, EvoEmbedding reranking, and the Atomic prior. Data, K, source-turn serialization, Answer model, Answer prompt, Judge model, and Judge prompt remain fixed within each benchmark.

Table 7: Controller-off paired answer accuracy (%) by answerability. N is the cohort size; Δ pp is the complete-path minus Direct+Evo accuracy difference in percentage points. Exact p values use a two-sided McNemar test.

Benchmark	Cohort	N	Direct+Evo	Complete path	Δ pp / p
LoCoMo	All	1,986	72.16	**75.13**	+2.97 / 6.40×10^{-5}
	Answerable	1,540	85.00	**87.73**	+2.73 / 4.86×10^{-4}
	Unanswerable	446	27.80	**31.61**	+3.81 / 0.060
LongMemEval-S	All	500	82.80	**84.40**	+1.60 / 0.302
	Answerable	470	84.04	**86.38**	+2.34 / 0.117
	Unanswerable	30	**63.33**	53.33	−10.00 / 0.375

Paired percentile bootstrap intervals over complete histories are $[+1.21, +4.72]$ points for LoCoMo and $[-1.00, +4.20]$ points for LongMemEval-S, using 20,000 resamples and seed 20260916. They condition on observed outputs and do not include repeated-generation or repeated-Judge variance.

The LoCoMo evidence-path comparison uses one fixed component ordering for both arms. Relative to the earlier refusal-ablation inventory, that ordering changes 27 serialized Answer payloads in

the complete arm; those payloads were executed again and the other 1,959 were reused by exact request identity. The resulting 1,492/1,986 controller-off score is therefore the baseline for this paired evidence-path comparison. The separate refusal ablation retains its own fixed 1,486/1,986 controller-off baseline and is interpreted only within that pair.

B.3 Category and evidence-conditioned results

Table 8: Complete category accuracy (%). Correct gives correct/total questions; category accuracy uses the benchmark-specific judge.

Benchmark/category	Correct	Accuracy
LongMemEval-S: single-session user	70/70	100.00
LongMemEval-S: multi-session	101/133	75.94
LongMemEval-S: preference	16/30	53.33
LongMemEval-S: temporal	113/133	84.96
LongMemEval-S: knowledge update	73/78	93.59
LongMemEval-S: assistant	54/56	96.43
LongMemEval-S overall	427/500	85.40
LoCoMo: single-hop	287/321	89.41
LoCoMo: multi-hop	235/282	83.33
LoCoMo: temporal	59/96	61.46
LoCoMo: open-domain	763/841	90.73
LoCoMo: adversarial	302/446	67.71
LoCoMo overall	1,646/1,986	82.88

Table 9: Answer correctness conditioned on complete annotated source coverage. Counts cover answerable questions with valid source annotations.

Benchmark	Evidence state	Correct	Wrong	Total
LongMemEval-S	all sources covered	386	46	432
LongMemEval-S	at least one missing	20	18	38
LoCoMo	all sources covered	1,152	109	1,261
LoCoMo	at least one missing	190	84	274

Five answerable LoCoMo questions without usable source annotations (two correct and three incorrect) are omitted. Six of eight LongMemEval-S final false refusals and 26 of 41 LoCoMo final false refusals occur under complete annotated coverage.

Table 10: Controller and final-output refusal confusion matrices. Refusal is the positive class; TP/FP/FN/TN denote true positives, false positives, false negatives, and true negatives. Signals count forced controller refusals, while Final refusals includes Answer Agent abstentions.

Benchmark	Signals	Final refusals	Controller TP/FP/FN/TN	Final TP/FP/FN/TN
LongMemEval-S	5	27	5/0/25/470	19/8/11/462
LoCoMo	221	323	209/12/237/1,528	282/41/164/1,499

The LoCoMo candidate generator proposes 332 cases (130 deterministic and 202 semantic-fallback); the verifier retains 221 forced refusals. The 102 other final refusals arise from the selected Answer outputs. LongMemEval-S has 22 pre-existing Answer-path refusals; the five forced refusals yield 27 final refusals. Controller and final-output metrics measure different decisions.

B.4 PAIRED REFUSAL RESULTS

The LoCoMo ablation fixes retrieval results, primary evidence, answer model, prompts, and Judge. Relative to no refusal, the verifier yields 167 wrong-to-correct and seven correct-to-wrong changes, a net gain of 160. The exact paired McNemar p-value is 7.39×10^{-41}.

Table 11: LongMemEval-S source-grounded verifier. Accuracy is in %; Gains and Losses count paired wrong-to-correct and correct-to-wrong changes.

Policy	Correct	Accuracy	Gains	Losses
Without forced refusal	422/500	84.40	–	–
With source-grounded verifier	**427/500**	**85.40**	5	0

C PROMPTS AND INTERFACES

C.1 ATOMIC-WRITING CONTRACT

The writer is instructed to create independently retrievable facets for durable facts, events, preferences, relationships, plans, quantities, dates, and constraints; preserve attribution, value, unit, time, negation, uncertainty, and event state; normalize time only when unambiguous; avoid unstated identity or causal inferences; and attach every memory to exact source-turn identifiers. Each stored item contains a stable ID, text, source-turn IDs, mention time, time expression, retrieval tags, typed entities, and a structured claim with subject, predicate, object, attribute, time, and polarity. Generated text is a retrieval index; final answers are grounded in the linked raw turns.

C.2 SHARED ANSWER RUNTIME AND BENCHMARK ADAPTERS

Both benchmarks use the same Answer Agent implementation, response schema, and source-turn payload formatter. Their common instruction core requires the model to answer only from selected evidence; identify the requested subject, relation, value, qualifier, scope, and unit; avoid substituting nearby roles or event states; choose and execute the requested lookup, aggregation, temporal, update, or preference operation; preserve units; and distinguish completed events from plans and recommendations.

The common logical request contains `question`, a task contract, evidence coverage, a source-linked evidence record, and evidence units. LongMemEval-S additionally serializes question-level `current_time`; LoCoMo has no equivalent question-time field, although its source turns retain session timestamps. Each evidence unit has a stable source-turn identifier and complete timestamped source text. The JSON response contains evidence status, operation, compact derivation, reason, final answer, and supporting IDs.

The fixed system prompts are benchmark-specific and remain identical between paired retrieval arms within a benchmark. LongMemEval-S adds explicit source-turn execution instructions for deterministic time constraints, operand enumeration, state ordering, descriptive referents, and personalized advice. LoCoMo uses the shared instruction core without this additional block. These prompt differences are held fixed in every reported within-benchmark ablation.

C.3 REFUSAL INTERFACES

The LoCoMo fallback marker receives the question frame and exactly three Atomic candidates. It may label each candidate `DIRECT_SUPPORT`, `PARTIAL_SUPPORT`, `ROLE_MISMATCH`, `ATTRIBUTE_MISMATCH`, `RELATION_MISMATCH`, or `IRRELEVANT`. It must not answer, retrieve, or infer that the full store lacks an answer. Any support label vetoes refusal and malformed output fails open.

The LongMemEval-S adapter receives the question, compact frame, current time, and Atomic evidence window. Its output records support status, a typed conflict field named `certificate`, and cited memories. The shared verifier reads the question, task contract, draft answer, candidate certificate, and complete restored turns. LoCoMo role-substitution certificates are normalized

to TARGET_FRAME_SUBSTITUTION before this call. Only a confirmed target-frame conflict without support can force refusal; plan-as-completed, advice/preference, missing operand, unresolved, and invalid cases fail open. The verifier prompt and hard-decision rule were developed on LongMemEval-S and transferred to LoCoMo without label-based adjustment in this experiment.

C.4 JUDGE PROMPTS

We reproduce the English Judge templates used in our evaluation. LoCoMo uses Mem0's accuracy prompt (Chhikara et al., 2025) (source code), sent as a single user message. LongMemEval-S uses the official question-type-specific prompts (Wu et al., 2025) (source code); the system instruction used by our evaluator is also reproduced below. Reference answers and question-type or abstention annotations are used only for post-generation scoring and are not provided to the memory system or Answer Agent. Braced fields denote inputs substituted for each evaluated response. Only display line wrapping is adjusted.

The general template applies to `single-session-user`, `single-session-assistant`, and `multi-session` questions. Temporal-reasoning, knowledge-update, and single-session-preference questions use their respective templates. Unanswerable items use the abstention template, which takes precedence over the question type.

LoCoMo user prompt.

```
Your task is to label an answer to a question as 'CORRECT'
or 'WRONG'. You will be given the following data:
(1) a question (posed by one user to another user),
(2) a 'gold' (ground truth) answer,
(3) a generated answer
which you will score as CORRECT/WRONG.
The point of the question is to ask about something one user
should know about the other user based on their prior
conversations.
The gold answer will usually be a concise and short answer
that includes the referenced topic, for example:
Question:  Do you remember what I got the last time I went
to Hawaii?
Gold answer:  A shell necklace
The generated answer might be much longer, but you should be
generous with your grading - as long as it touches on the
same topic as the gold answer, it should be counted as
CORRECT.
For time related questions, the gold answer will be a
specific date, month, year, etc.  The generated answer might
be much longer or use relative time references (like "last
Tuesday" or "next month"), but you should be generous with
your grading - as long as it refers to the same date or time
period as the gold answer, it should be counted as CORRECT.
Even if the format differs (e.g., "May 7th" vs "7 May"),
consider it CORRECT if it's the same date.
Now it's time for the real question:
Question:  {question}
Gold answer:  {gold_answer}
Generated answer:  {generated_answer}
First, provide a short (one sentence) explanation of your
reasoning, then finish with CORRECT or WRONG.
Do NOT include both CORRECT and WRONG in your response, or
it will break the evaluation script.
Just return the label CORRECT or WRONG in a json format with
the key as "label".
```

LongMemEval-S system instruction.

```
You are an impartial LongMemEval answer evaluator. Follow
the user prompt exactly and return yes or no only.
```

LongMemEval-S general questions.

```
I will give you a question, a correct answer, and a response
from a model. Please answer yes if the response contains
the correct answer. Otherwise, answer no. If the response
is equivalent to the correct answer or contains all the
intermediate steps to get the correct answer, you should
also answer yes. If the response only contains a subset of
the information required by the answer, answer no.
Question: {question}
Correct Answer: {answer}
Model Response: {response}
Is the model response correct? Answer yes or no only.
```

LongMemEval-S temporal reasoning.

```
I will give you a question, a correct answer, and a response
from a model. Please answer yes if the response contains
the correct answer. Otherwise, answer no. If the response
is equivalent to the correct answer or contains all the
intermediate steps to get the correct answer, you should
also answer yes. If the response only contains a subset of
the information required by the answer, answer no. In
addition, do not penalize off-by-one errors for the number
of days. If the question asks for the number of
days/weeks/months, etc., and the model makes off-by-one
errors (e.g., predicting 19 days when the answer is 18), the
model's response is still correct.
Question: {question}
Correct Answer: {answer}
Model Response: {response}
Is the model response correct? Answer yes or no only.
```

LongMemEval-S knowledge updates.

```
I will give you a question, a correct answer, and a response
from a model. Please answer yes if the response contains
the correct answer. Otherwise, answer no. If the response
contains some previous information along with an updated
answer, the response should be considered as correct as long
as the updated answer is the required answer.
Question: {question}
Correct Answer: {answer}
Model Response: {response}
Is the model response correct? Answer yes or no only.
```

LongMemEval-S user preferences.

```
I will give you a question, a rubric for desired
personalized response, and a response from a model. Please
answer yes if the response satisfies the desired response.
Otherwise, answer no. The model does not need to reflect
all the points in the rubric. The response is correct as
long as it recalls and utilizes the user's personal
information correctly.
Question: {question}
Rubric: {answer}
Model Response: {response}
Is the model response correct? Answer yes or no only.
```

LongMemEval-S unanswerable questions.

```
I will give you an unanswerable question, an explanation,
and a response from a model.  Please answer yes if the model
correctly identifies the question as unanswerable.  The
model could say that the information is incomplete, or some
other information is given but the asked information is not.
Question:  {question}
Explanation:  {answer}
Model Response:  {response}
Does the model correctly identify the question as
unanswerable?  Answer yes or no only.
```

C.5 TWO-ANNOTATOR HUMAN AUDIT OF THE LLM JUDGES

Two human annotators independently reviewed the same 200 frozen GroundedMem responses in different row orders. We sampled 100 responses per benchmark, with 50 Judge-`CORRECT` and 50 Judge-`WRONG` decisions, and allocated cases proportionally across the benchmark's question categories. Each annotator saw only the question, reference answer, system answer, and applicable benchmark rubric; Judge decisions, question identifiers, output provenance, sampling strata, and the other annotator's labels were hidden until both annotation files were frozen.

Because Judge labels were deliberately balanced, raw sample agreement is a stratified diagnostic rather than a prevalence estimate. We therefore also report an estimate that weights agreement conditional on each Judge label by that label's frequency in the complete paper-facing cohort. Wilson intervals apply to raw agreement; the weighted estimate uses 20,000 stratified bootstrap resamples with seed 20260925.

Table 12: Two-annotator human audit of the benchmark-specific LLM Judges. Panel (a) reports pre-adjudication human–human reliability. Panel (b) reports each annotator's agreement with the Judge. J-C/H-W counts Judge-`CORRECT` versus human-`WRONG`; J-W/H-C is the reverse. Weighted agreement restores full-cohort Judge-label prevalence.

(a) Human–human agreement

Benchmark	N	Agreement [95% CI]	κ	Disagreements
LongMemEval-S	100	100.0 [96.3, 100.0]	1.000	0
LoCoMo	100	98.0 [93.0, 99.4]	0.960	2
Pooled	200	99.0 [96.4, 99.7]	0.980	2

(b) Judge–human agreement

Benchmark	Annot.	N	Agreement [95% CI]	κ	J-C/H-W / J-W/H-C	Weighted
LongMemEval-S	A	100	94.0 [87.5, 97.2]	0.880	0 / 6	98.2
LongMemEval-S	B	100	94.0 [87.5, 97.2]	0.880	0 / 6	98.2
LoCoMo	A	100	97.0 [91.5, 99.0]	0.940	1 / 2	97.7
LoCoMo	B	100	97.0 [91.5, 99.0]	0.940	2 / 1	96.3

Before adjudication, the annotators agreed on 198/200 cases (99.0%; $\kappa = 0.980$): 100/100 on LongMemEval-S and 98/100 on LoCoMo. The two LoCoMo disagreements concern whether partial coverage or an implied social relation is sufficient under the benchmark's generous rubric. We retain both labels rather than select a post-hoc gold label. Each annotator agreed with the Judge on 191/200 cases (95.5%; $\kappa = 0.910$). Both identified the same six LongMemEval-S preference cases as `CORRECT` where the Judge returned `WRONG`. On LoCoMo, both achieved 97.0% raw agreement, while the class-prevalence-weighted estimates differ because the two human disagreements fall in different Judge-label strata.

D Implementation and Evaluation Details

D.1 System configuration

Table 13: System settings. RRF is reciprocal rank fusion; Top-K is the number of retained candidates or source turns at that stage.

Component	Setting
Atomic writer	DeepSeek-V4-Flash; temperature 0; at most 8,192 output tokens
Atomic route	Top-30 Atomic memories
Direct route	BM25 Top-100 plus Qwen3-Embedding-0.6B Top-100
Direct/cross-view fusion	Equal-weight RRF, constant 60, Top-100
Contextual reranker	EvoEmbedding-4B over Top-100
Final Atomic prior	Weight 0.5, RRF constant 10
Answer evidence	LongMemEval-S Top-10; LoCoMo Top-15 complete turns
Answer generation	GPT-4.1-mini; temperature 0; at most 640 tokens; reasoning disabled
Answer adapter	Shared runtime and source-turn schema; benchmark-specific prompt; question time available only on LongMemEval-S
Refusal audit	Benchmark candidate adapters, then shared source-grounded verifier on draft and restored turns
Evaluation	Benchmark-specific Judge prompt, parsing, and metric rules

Retrieval models run locally. Answer generation, judging, Atomic writing, and semantic checks use external APIs.

Each conversation is split into session-bounded episodes of at most ten messages, with two-message overlap and at most 1,024 tokens per episode. The writer processes batches of eight target turns with up to two preceding turns as context. At temperature zero, it emits at most two turn-local facets per substantive turn and at most two local aggregate facets; an aggregate cites at most three turns. Every stored item must include source-turn identifiers. Both benchmarks use the same writer configuration. Source adapters bind stored items to coordinates in the original history, and the resulting mappings are fixed before evaluation. Benchmark labels are excluded from memory writing.

Table 14: Memory-writing statistics. Episodes are bounded history segments; Writer requests count calls to the Atomic-memory writer.

Benchmark	Episodes	Atomic memories	Writer requests
LongMemEval-S	23,867	212,206	42,367
LoCoMo	272	4,270	848

The question, reference answer, answerability label, and annotated gold turns are not inputs to memory writing.

Shared refusal design and transfer. Each benchmark first proposes possible conflicts from bounded Atomic evidence. After the Answer Agent drafts a response from complete source turns, the same source-grounded verifier checks the question, task contract, draft, candidate certificate, and restored turns. Its prompt and decision rule were developed on LongMemEval-S and transferred to LoCoMo after certificate normalization. Positive source support vetoes refusal; missing or unresolved evidence retains the draft. Only a verified `TARGET_FRAME_SUBSTITUTION` without support may force canonical abstention. `PLAN_AS_COMPLETED` remains a non-enforcing signal; advice, preference, recommendation, and malformed cases fail open.

D.2 LABEL ISOLATION

Table 15: Information visible to each stage at runtime or evaluation. Gold means benchmark-provided reference information; "yes" and "no" indicate whether a stage can access the field.

Stage	Question	History	Gold answer	Labels	Gold turns
Memory writing	no	yes	no	no	no
Retrieval/ranking	yes	yes	no	no	no
Candidate adapter	yes	Atomic window	no	no	no
Source-grounded verifier	yes	selected turns	no	no	no
Answer agent	yes	selected turns	no	no	no
Evaluation	yes	protocol-specific	yes	yes	yes

Memory writing uses raw histories; retrieval and control also use the question and use question time only when the benchmark provides it. Answer generation receives selected source turns. Reference answers, benchmark categories, answerability labels, and gold sources are reserved for evaluation.

D.3 EFFICIENCY AND COST ACCOUNTING

Following REMem (Shu et al., 2026), we separate indexing and inference in time/memory and token/cost tables. Local retrieval and remote LLM calls were not synchronized, so no end-to-end latency is constructed. Hardware is a Xeon Gold 6226R, 60 GiB RAM, and one 24 GiB RTX 4090.

Time and memory. Index time adds the concurrent Atomic-writer span and cold index build. Local inference/Q is single-GPU Direct/fusion/Evo batch time per question, including switching/serialization but excluding Atomic retrieval and remote LLM calls. Peak host memory is maximum process-tree RSS in a fresh zero-API sample; GPU memory lacked reliable process attribution and is omitted.

Table 16: REMem-style running time and host-memory usage. Local inference/Q is throughput-equivalent, not online end-to-end latency.

Benchmark	Index time (s)	Local inference/Q (s)	Max host memory (GiB)
LoCoMo	580	5.76	17.81
LongMemEval-S	51,245	14.13	20.98

Persistent footprints are 87.1 MiB/5.92 GiB; peak-memory samples cover 10/50 questions. Hardware and service boundaries preclude direct comparison with REMem.

Token usage and cost. Indexing reconstructs frozen Atomic-writer requests with `o200k_base` plus 11 tokens/request; inference combines Answer and controller usage and excludes Judge calls. Cost is an Answer-only lower bound at GPT-4.1-mini's $0.40/M input and $1.60/M output rates (September 25, 2026; source); other prices were not retained. Indexing/Answer counts are proxies; controller counts are provider usage including reasoning.

Table 17: REMem-style frozen-cohort token usage and available cost (millions).

	Indexing phase		Inference phase			
Benchmark	Input	Output	Input	Output	Total	Cost lower bound (USD)
LoCoMo	3.00	0.58	19.61	3.47	26.65	7.37
LongMemEval-S	238.84	34.81	8.99	1.25	283.89	2.60