CAUSEWAY: AUDITABLE CAUSAL MEMORY FOR LONG-HORIZON LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent memory retrieves text that resembles the query, but an agent needs evidence that supports and influences the answer. Similarity and graph proximity can surface evidence that is stale, redundant, or confounded by a shared upstream cause, while retrieval alone does not establish whether an answer depends on that evidence. Destructive consolidation compounds this problem by deleting the history that temporal questions depend on. We present CAUSEWAY, a memory substrate that connects causal structure, evidence influence, and source auditing. CAUSEWAY compiles interaction history into a bitemporal causal graph whose typed edges become load-bearing after passing precedence, mechanism, and evidence gates. On its interventional path, each query is compiled into an explicit answer variable, and candidates are scored by contrasting their presence and absence in reader-visible evidence, with graph-selected background and a negative-control penalty. On the graph path, a source-linked proof supports answer generation and applicable removal/polarity checks; failed support checks trigger source replay or abstention. Non-destructive writes preserve earlier states for historical queries. Instantiated training-free with MiMo v2.5, CAUSEWAY achieves the highest Overall scores among evaluated methods on LoCoMo, LoCoMo-Plus, and MemoryAgentBench. Its strongest gains concern evidence localization, evolving states, and persistent constraints. Matched LoCoMo comparisons show that verification improves Answerable F1 by 6.30 points while reducing coverage by 1.29 points. On LoCoMo-Plus, targeted evidence use improves accuracy over Full Context with 76.99% fewer query-stage LLM tokens. Our code is available at `https://anonymous.4open.science/r/Causeway-3F66/`.

1 INTRODUCTION

Long-horizon agents are judged by what they accumulate. An assistant tracks a user's changing job, diet, and relationships over months; a coding agent revisits a repository across dozens of sessions; a research copilot must recall why an earlier decision was made. Context windows have grown, but interaction histories grow faster and more redundantly, so what turns a stateless predictor into an agent is memory: machinery that writes experience out, organizes it, and reads part of it back. The prevailing design is external and textual: MemGPT pages memory like an operating system, generative agents and Reflexion add reflective summaries and verbal feedback, Mem0 consolidates salient facts, and A-MEM, Zep, and MAGMA impose graph structure over notes, entities, and temporal relations Packer et al. (2023); Park et al. (2023); Shinn et al. (2023); Chhikara et al. (2025); Xu et al. (2026); Rasmussen et al. (2025); Jiang et al. (2026); Yang et al. (2026). LoCoMo, LongMemEval, and LoCoMo-Plus have made this pipeline measurable and have driven steady gains on multi-session recall Maharana et al. (2024); Wu et al. (2024); Li et al. (2026). Yet stronger recall does not translate uniformly into better use of historical constraints: across methods, performance on LoCoMo and LoCoMo-Plus can diverge substantially (Figure 1).

This divergence points to a deeper problem: these systems optimize a proxy. They rank memories by resemblance to the query, while the agent needs memories that determine the answer, and the two objectives come apart in three ways that matter. First, similarity is blind to influence. A memory can be topically perfect yet stale, redundant, or misleading, and it can look decisive only because it shares an upstream cause with the answer: a looming deadline drives both *ordered more coffee* and *slept badly*, so the coffee appears causal. Every such passage further dilutes attention and invites

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.

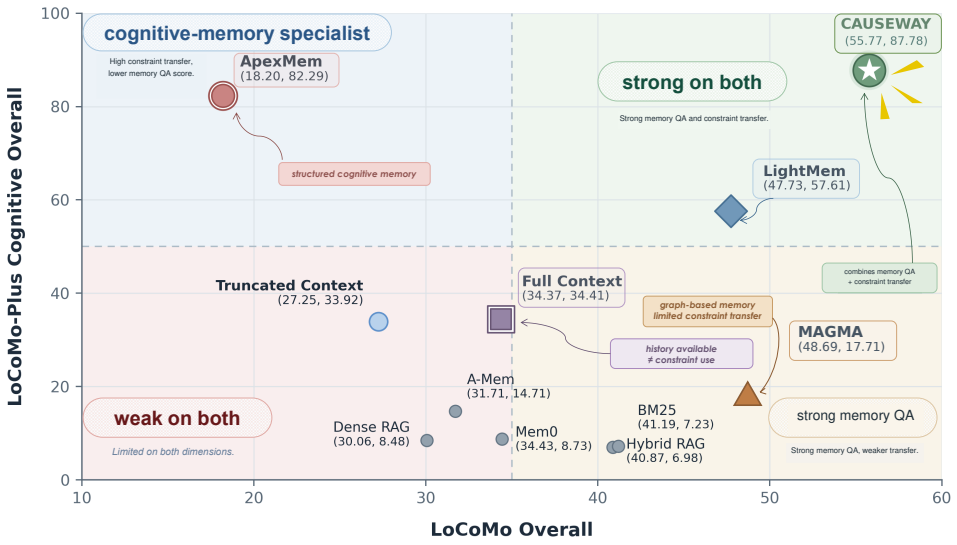

Figure 1: LoCoMo Overall versus LoCoMo-Plus Cognitive Overall for the methods in Table 1.

lost-in-the-middle errors Liu et al. (2024); Srivastava (2026). Second, support is asserted rather than checked. A memory system presents its evidence as the basis for the output but does not test the claim, so an answer that would be produced identically with the evidence deleted is hard to distinguish from one that genuinely depends on it. Third, memory is written destructively. Consolidation overwrites superseded facts and quietly removes the record that temporal and knowledge-update questions interrogate, and when the surviving memory is insufficient the system still answers. Repairing these failures is hard, because dialogue arrives without causal annotation, interventional tests are expensive at retrieval time, shared-cause evidence must be handled without a known graph, and abstention must be integrated with evidence checks.

Recent work moves toward influence without fully arriving. Graph memories add relational scaffolding and bitemporal edges, yet retrieval still follows proximity or reachability, which favors what co-occurred over what mattered Xu et al. (2026); Rasmussen et al. (2025); Jiang et al. (2026); Yang et al. (2026). Bitemporal operator algebras such as TOKI focus on contradiction resolution over versioned stores Wang (2026). CMI comes closest, selecting memories by whether their presence, absence, or perturbation changes the model's answer, but it operates over a flat bank without graph-selected common-cause context, temporal versioning, or an audit of the answer it finally emits Srivastava (2026). ActMem and REMI build causal-semantic graphs and reason counterfactually, but emphasize constraint inference and factor criticality rather than their integration with quantified evidence sensitivity, answer auditing, and temporal versioning Zhang et al. (2026); Raman et al. (2025). The dependency edges of ContextWeaver encode tool-output plumbing rather than Pearlian causation Wu et al. (2026). Our design brings together graph-conditioned evidence-intervention scoring, tests of answer dependence on evidence, queryable historical states, and selective answering.

We introduce CAUSEWAY, which reorganizes memory around leverage instead of likeness. Its substrate is a bitemporal causal graph over source-grounded events. An edge becomes *load-bearing* only after clearing precedence, mechanism, and source-evidence gates, and otherwise survives as a queryable hypothesis, preserving the distinction between admitted relations and unconfirmed hypotheses. At query time CAUSEWAY first names what it must explain, compiling the question into an explicit outcome variable. Its interventional path contrasts candidate presence and absence while retaining graph-selected common-cause context, and penalizes influence on unrelated probes; these are confounding-inspired controls on evidence sensitivity Pearl (2009); Srivastava (2026). The reader receives a directed causal proof in place of a bag of snippets, and that proof is cross-examined through critical-node removal and applicable edge-polarity checks; a failed proof can trigger rederivation from the raw sources behind the disputed nodes Zhang et al. (2026); Raman et al. (2025). Writes are non-destructive, so valid time and transaction time keep the current state and historical causal states reconstructable, and the evidence interface supports answering, abstention, and source replay. The evaluated realization uses MiMo v2.5 without parameter updates; graph construction

is cached separately from query-time reading, and task adapters activate the relevant retrieval and verification operations (Appendix A.1).

Our contributions are the following. (i) **Memory as a bitemporal causal ledger**, in which mechanism-gated, provenance-anchored edges extend flat causal banks and semantic-causal knowledge graphs into a versioned causal substrate. (ii) **Graph-conditioned evidence interventions**, combining estimated answer sensitivity with causal paths, temporal validity, and non-specific-influence controls. (iii) **A counterfactual answer-dependence audit** that tests dependence on retrieved support, with source replay when a proof fails cross-examination. (iv) **Non-destructive causal versioning** that retains valid-time and recorded-time views for knowledge updates. (v) **Selective answering through evidence checks**, evaluated through both answer quality and response coverage. Against context, retrieval, and memory-system baselines on LoCoMo, LoCoMo-Plus, and MemoryAgentBench, CAUSEWAY improves Overall over LightMem by 8.05, 30.17, and 21.55 points, and component studies isolate retrieval-route replacement, temporal modeling, and verification on frozen subsets. What a long-horizon agent should remember is not what resembles the question but what the answer rests on.

2 RELATED WORK

Retrieval, summary, and graph memory. Agent memories store interactions externally and read them back through OS-style paging, reflection, or verbal feedback Packer et al. (2023); Park et al. (2023); Shinn et al. (2023), fact consolidation, linked notes, or temporal and graph structures Chhikara et al. (2025); Xu et al. (2026); Rasmussen et al. (2025); Jiang et al. (2026); Yang et al. (2026), with LightMem and ApexMem targeting efficiency and temporal reasoning Fang et al. (2026); Banerjee et al. (2026); they build on sparse, dense, and graph-based retrieval Robertson & Zaragoza (2009); Karpukhin et al. (2020); Lewis et al. (2020); Gutiérrez et al. (2024); Edge et al. (2024). CAUSEWAY keeps structured retrieval and provenance but ranks evidence by its influence on the answer Liu et al. (2024).

Causal and counterfactual memory. CMI selects memories by controlled interventions; ActMem and REMI reason counterfactually over causal-semantic graphs and personal causal schemas; ContextWeaver validates dependency traces with execution feedback Srivastava (2026); Zhang et al. (2026); Raman et al. (2025); Wu et al. (2026). CAUSEWAY unifies evidence sensitivity, answer-dependence auditing, bitemporal versioning, and source replay.

Temporal memory and benchmarks. Interval relations, situation calculus, and valid/transaction-time databases ground temporal memory, and TOKI resolves contradictions in bitemporal agent stores Allen (1983); McCarthy & Hayes (1981); Snodgrass & Ahn (1985); Wang (2026). Long-horizon memory suites Maharana et al. (2024); Wu et al. (2024); Li et al. (2026); Du et al. (2024); Hu et al. (2026) test cross-session recall and knowledge updates, complementing long-context benchmarks Bai et al. (2024); Hsieh et al. (2024); Zhang et al. (2024).

Verification and selective prediction. Self-RAG, CRAG, ERASER, and Chain-of-Verification check retrieval, rationales, and generated answers Asai et al. (2024); Yan et al. (2024); DeYoung et al. (2020); Dhuliawala et al. (2024); ProReason, TRACE, and LC-ERD structure or evaluate reasoning Zhou et al. (2025); Chen et al. (2026a;b). CAUSEWAY borrows intervention and adjustment ideas from causal inference Pearl (2009); unlike calibrated conformal abstention and risk control Yadkori et al. (2024); Angelopoulos et al. (2024), its evidence checks trigger abstention or source replay over versioned conversational memory, without coverage guarantees.

3 METHODOLOGY: CAUSEWAY

We propose **CAUSEWAY**, a *Temporal-causal Reasoning and Counterfactual Evidence Memory* for long-horizon LLM agents that represents history as a non-destructive temporal-causal graph and retrieves evidence by estimated *influence on the reader*, causal paths, and temporal validity. CAUSEWAY integrates intervention-based memory selection Srivastava (2026), causal-semantic graphs with counterfactual reasoning Zhang et al. (2026); Raman et al. (2025), and dependency-structured traces with execution feedback Wu et al. (2026) into a single versioned, abstention-aware framework, instantiated without parameter updates. Figure 2 summarizes the pipeline.

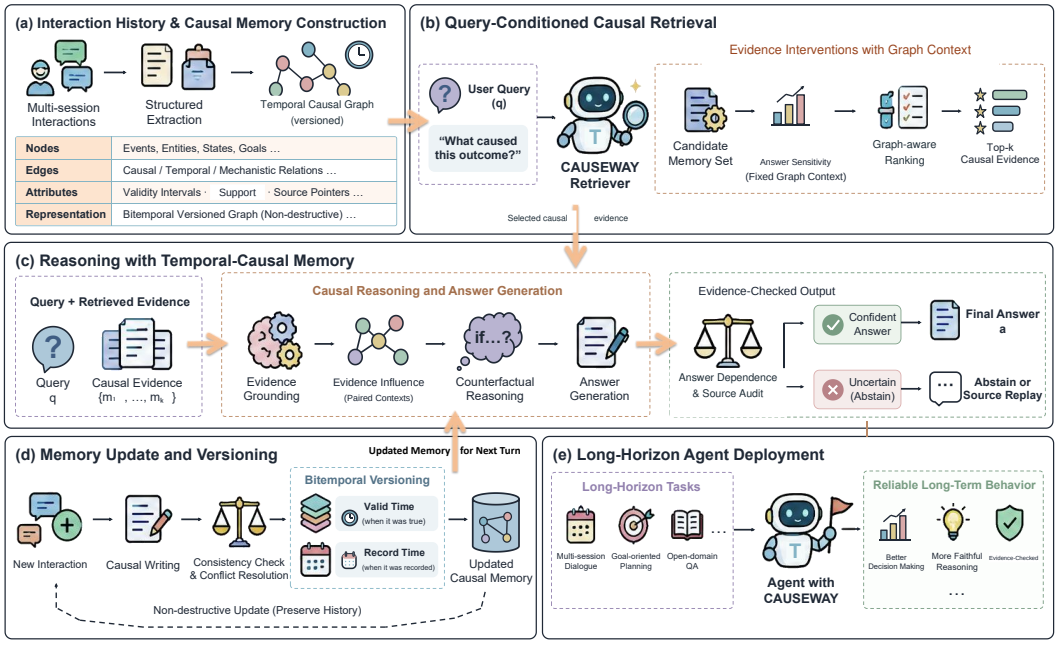

Figure 2: Overview of CAUSEWAY: source-linked temporal-causal memory, query-specific evidence selection, post-generation dependence/source auditing, and non-destructive updates. Failed checks lead to source replay or abstention.

3.1 FROM RESEMBLANCE TO INFLUENCE: PROBLEM FORMULATION

Given history $\mathcal{H} = \{m_1, \ldots, m_T\}$, query q, and reader p_θ, conventional retrieval selects context $C_q \subset \mathcal{H}$ by similarity and answers with $p_\theta(y \mid q, C_q)$ Srivastava (2026); Liu et al. (2024). CAUSEWAY instead maintains a causal memory state $\mathcal{G}_T = (V_T, E_T)$ whose nodes are temporally grounded events and whose edges encode causal, temporal, update, support, and conflict relations, compiles the query into an outcome variable Y_q, and returns a causal proof subgraph $P_q \subseteq \mathcal{G}_T$, a source-linked dependency structure selected for evidence influence:

$$P_q^\star = \arg\max_{P \subseteq \mathcal{G}_T} \; \mathrm{CE}(P \to Y_q) - \lambda\, c(P) - \mu\, \mathrm{Conflict}(P),$$

where CE is an estimated evidence-sensitivity score, $c(P)$ is token/latency cost, and $\mathrm{Conflict}(P)$ penalizes contradictions. Influence is unsigned: high CE alone establishes neither correctness nor completeness. Proof-level aggregation and its coverage limits are specified in Appendix A.2.

3.2 HISTORY AS A BITEMPORAL CAUSAL GRAPH: THE VERSIONED LEDGER

CAUSEWAY stores memory as $\mathcal{G}_T = (V_T, E_T, \mathcal{S})$, with event set V_T, relation set E_T, and an append-only source log $\mathcal{S}$ that anchors every node and edge to raw messages for audit and replay. Each event carries content, type, entities, a validity interval, confidence, and provenance, and each edge additionally records a relation type, causal polarity $\sigma \in \{-1, +1\}$, mechanism, temporal lag ℓ, and its own validity interval:

$$v = \big(x_v, \mathrm{type}_v, E_v, [t_v^{\mathrm{start}}, t_v^{\mathrm{end}}), \mathrm{conf}_v, \mathrm{src}(v)\big),$$
$$e_{u \to v} = \big(u, v, r, \sigma, \ell, [t_e^{\mathrm{start}}, t_e^{\mathrm{end}}), \mathrm{conf}_e, \mathrm{src}(e)\big),$$

with node types $\mathrm{type}_v \in \{\text{fact, pref, action, obs, decision, constraint, goal, outcome}\}$ and relation types $r \in \{\texttt{causes}, \texttt{enables}, \texttt{prevents}, \texttt{updates}, \texttt{contradicts}, \texttt{supports}, \texttt{before}, \texttt{after}\}$, so that the same substrate expresses semantic, temporal, and mechanistic structure that flat or purely associative stores cannot Jiang et al. (2026); Yang et al. (2026). Appendix A.10 and Figure 6 give source-level cases of version-aware retrieval, temporal grounding, and goal–event links.

3.3 Writing Mechanisms, Not Just Facts: Event and Edge Extraction

A model-backed *Writer* first extracts candidate event nodes from each interaction segment, preserving exact dates, quantities, negations, and updates, and emits source-grounded JSON via constrained tool calls so that extraction is inspectable and reversible.

An *Edge Miner* then proposes relations, but an edge is admitted as an *active*, source-supported causal claim only if it passes three gates rather than mere topical co-occurrence: temporal precedence $\mathrm{Temporal}(u,v) = \mathbf{1}[t_u < t_v]$, a source-linked proposed mechanism $\mathrm{Mechanism}(u,v) = \mathbf{1}[\text{source-grounded mechanism}]$, and source evidence $\mathrm{Evidence}(u,v) = \mathbf{1}[\mathrm{src}(u) \cup \mathrm{src}(v) \text{ supports } e]$. The composite admission score multiplies the three gate confidences, $\mathrm{conf}_e = \mathrm{conf}_{\mathrm{temp}} \cdot \mathrm{conf}_{\mathrm{mech}} \cdot \mathrm{conf}_{\mathrm{evid}}$, and edges below threshold are retained as *hypotheses* (queryable but never load-bearing in a proof), keeping uncertain relations available without treating them as established support Zhang et al. (2026). These are admission scores, not calibrated probabilities; operational rules appear in Appendix A.2.

3.4 Naming the Unknown: Query-to-Outcome Compilation

An *Outcome Compiler* maps a query q to an explicit answer variable $Y_q = f_{\mathrm{out}}(q)$. For example, "Why did the user stop using DoorDash?" asks for the reason for stopping; "What is the user's current workplace?" asks for the current employer. The compiler also specifies the relevant entities, query time, required relations, answer format, and whether counterfactual checking is needed; together, these constraints $\mathcal{C}_q$ determine which graph region to retrieve and how the answer should be expressed.

3.5 Retrieval by Intervention, Not Resemblance: Interventional Scoring

For candidate evidence Z, let I_Z denote its visibility to the reader. CAUSEWAY contrasts model-estimated answer distributions under $I_Z = 1$ and $I_Z = 0$, retaining graph-selected background $A_q(Z)$:

$$\mathrm{CE}(Z; Y_q) = D_{\mathrm{TV}}\Big(\widehat{p}_\theta(Y_q \mid I_Z = 1, q, A_q(Z)),\ \widehat{p}_\theta(Y_q \mid I_Z = 0, q, A_q(Z))\Big),$$

where $\widehat{p}_\theta$ is reported by one joint prompt containing both evidence worlds; TV is computed in code after label normalization. This extends CMI's flat-bank contrast to graph-conditioned evidence Srivastava (2026). The intervention changes reader-visible evidence, not historical events. The score is a model-estimated sensitivity proxy, not independently sampled reader probabilities or an identified real-world causal effect; the actual label rules and prompt contract appear in Appendix A.2.

The background rule selects represented common parents of Z and query-matched historical outcome proxies O_q, excluding Z, O_q, and represented descendants of Z. Thus $a \to Y_q$ abbreviates $a \to o$ for some $o \in O_q$, not an extracted edge into a new answer variable. This backdoor-inspired context rule is not a general backdoor identification criterion Pearl (2009). The final score blends evidence sensitivity with graph reachability, temporal validity, confidence, and a negative-control penalty for nodes that influence unrelated outcomes,

$$s(Z,q) = \alpha_1 \mathrm{CE}(Z;Y_q) + \alpha_2 \mathrm{PathRel}(Z,Y_q) + \alpha_3 \mathrm{TempValid}(Z,q) + \alpha_4 \mathrm{conf}(Z) - \alpha_5 \mathrm{Spurious}(Z),$$

where $\mathrm{Spurious}(Z)$ is high when Z changes the answer to semantically unrelated probe outcomes, a signal of non-specific prompt influence.

3.6 Chains over Bags: Causal Proof Assembly

Selected nodes are assembled into a directed acyclic proof $P_q = (V_q, E_q) \subseteq \mathcal{G}_T$ subject to temporal validity for every node, a confidence floor for every edge, and source-grounded reachability of every answer claim:

$$\forall v \in V_q,\ \mathrm{TempValid}(v,q) = 1; \qquad \forall e \in E_q,\ \mathrm{conf}_e \geq \eta; \qquad \exists\, \pi :\ v_{\mathrm{source}} \rightsquigarrow Y_q.$$

Here temporal validity follows the query's relevant time scope; an earlier cause need not remain current. Claim support does not ensure coverage of every requested item. The proof is serialized as an explicit causal chain `cause` → `intermediate state` → `outcome` and handed to the reader as the evidence for the graph-path reader, making its support legible. Task adapters and fallback reading are specified in Appendix A.1.

3.7 WOULD THE ANSWER SURVIVE?: COUNTERFACTUAL AUDITING

After graph-path generation, an *Auditor* tests local answer dependence on its proof: for each critical node $z \in P_q$ it recomputes the answer with z removed and marks z *locally influential* if the answer changes,

$$\hat{y}^{-z} = \arg\max_y p_\theta(y \mid q,\ P_q \setminus \{z\}), \qquad z \text{ locally influential} \iff \hat{y}^{-z} \neq \hat{y},$$

and, where applicable, it additionally tests edge-polarity reversal, $\hat{y}^{\mathrm{flip}(e)} = \arg\max_y p_\theta(y \mid q,\ P_q^{\mathrm{flip}(e)})$. An unchanged answer can reflect redundant support; a changed answer need not be correct. These tests complement source checks and do not certify proof completeness. Weak, unsupported, or contradictory proofs trigger *source replay*—retrieving the raw messages behind the disputed nodes and regenerating from explicit evidence Raman et al. (2025). If source support remains insufficient within the query budget, CAUSEWAY abstains.

3.8 THE PAST STAYS QUERYABLE: NON-DESTRUCTIVE CAUSAL VERSIONING

CAUSEWAY never overwrites; when new evidence supersedes a node or edge, the old element's valid-time is closed and a new version opened, $[t^{\mathrm{start}}_{\mathrm{old}}, t^{\mathrm{end}}_{\mathrm{old}}) = [t^{\mathrm{start}}_{\mathrm{old}}, t_{\mathrm{new}})$ and $[t^{\mathrm{start}}_{\mathrm{new}}, t^{\mathrm{end}}_{\mathrm{new}}) = [t_{\mathrm{new}}, \infty)$, while transaction time is recorded so that prior beliefs remain reconstructable. The ledger stores explicit version relations $\texttt{updates}(v_{\mathrm{new}}, v_{\mathrm{old}})$ and $\texttt{contradicts}(v_i, v_j)$, so "current" queries activate the latest valid causal state through the time filter while "historical" queries recover earlier states, preserving the distinction between when a fact holds and when it became available Wu et al. (2024); Li et al. (2026).

3.9 CACHED CONSTRUCTION AND TASK-SPECIFIC READING

The evaluated realization is training-free: query-time generation uses MiMo v2.5 at temperature zero with thinking disabled, and graph operations (`add_event_node`, `add_causal_edge`, `close_validity_interval`, `query_causal_graph`) are exposed as tool functions. Event and edge extraction persist source-grounded history in SQLite-backed caches, while the online graph path performs outcome compilation, candidate scoring, proof assembly, generation, and applicable auditing, with raw-source replay for unsupported proofs. Task adapters specialize this interface: Plus uses hierarchical session selection and MAB native-task readers, rather than uniformly executing CE and deletion. Appendix A.1 records route activation and post-audit fallbacks.

4 EXPERIMENTS

We evaluate whether temporal-causal memory improves the use of long interaction histories. Across three benchmarks, we examine overall performance, capability-specific gains, component effects, verification behavior, and query cost.

4.1 EXPERIMENTAL SETUP

Benchmarks and baselines. LoCoMo Maharana et al. (2024) covers single-hop, multi-hop, temporal, open-domain, and adversarial conversational questions; LoCoMo-Plus Li et al. (2026) tests whether earlier causal, state, goal, and value constraints inform responses to lexically distant queries; and MemoryAgentBench (MAB) Hu et al. (2026) covers accurate retrieval (AR), test-time learning (TTL), long-range understanding (LRU), and conflict resolution (CR). We compare context baselines (Full and Truncated Context), retrieval baselines (BM25, Dense RAG, and Hybrid RAG), and memory systems (A-Mem Xu et al. (2026), Mem0 Chhikara et al. (2025), LightMem Fang et al. (2026), MAGMA Jiang et al. (2026), and the paper-guided ApexMem* reimplementation Banerjee et al. (2026)). Methods share fixed evaluation sets: 600 LoCoMo questions, all 401 Cognitive questions, and 320 stratified MAB items. Component studies use matched subsets; protocols appear in Appendix A.1.

Metrics, generation, and comparisons. LoCoMo uses token F1 for answerable questions and the released refusal scorer for adversarial questions; Answerable F1 retains all answerable items, including abstentions, under the released scorer. LoCoMo-Plus uses binary accuracy under the released

Table 1: Performance across long-horizon memory benchmarks. Best scores are bold; CAUSEWAY is shaded. N/A denotes incomplete coverage. n gives the aggregation count (Appendix A.1).

Category / aggregate	n	A-Mem	Mem0	BM25	Light Mem	MAGMA	Apex Mem*	Dense RAG	Hybrid RAG	Trunc. ctx.	Full ctx.	Causeway
LoCoMo												
Overall	600	31.71	34.43	41.19	47.73	48.69	18.20	30.06	40.87	27.25	34.37	**55.77**
Ans. F1	465	14.89	16.04	35.08	48.25	38.74	22.41	15.13	34.88	13.87	22.63	**53.69**
Type macro	5	26.64	30.32	33.14	44.54	41.91	16.99	26.40	32.38	24.36	31.19	**50.83**
Multi-hop	85	11.50	13.20	22.50	**42.38**	23.22	23.45	11.05	21.65	18.20	21.92	37.39
Temporal	95	4.84	16.76	22.39	54.34	35.51	22.79	7.30	21.07	7.28	25.94	**57.38**
Open-domain	50	4.96	4.67	8.07	27.78	17.74	10.60	11.65	6.36	6.45	8.79	**34.19**
Single-hop	235	22.30	19.19	50.51	52.26	50.13	24.40	20.52	51.32	16.55	24.49	**62.24**
Adversarial	135	89.63	**97.78**	62.22	45.93	82.96	3.70	81.48	61.48	73.33	74.81	62.96
LoCoMo-Plus Cognitive												
Overall	401	14.71	8.73	7.23	57.61	17.71	82.29	8.48	6.98	33.92	34.41	**87.78**
Type macro	4	14.70	8.71	7.22	57.60	17.67	82.28	8.48	6.98	33.89	34.39	**87.76**
Causal	101	20.79	15.84	11.88	58.42	31.68	87.13	7.92	8.91	42.57	43.56	**94.06**
State	100	6.00	8.00	6.00	67.00	15.00	84.00	11.00	8.00	35.00	33.00	**86.00**
Goal	100	15.00	5.00	3.00	50.00	12.00	79.00	3.00	3.00	32.00	34.00	**82.00**
Value	100	17.00	6.00	8.00	55.00	12.00	79.00	12.00	8.00	26.00	27.00	**89.00**
MemoryAgentBench												
Overall	320	20.26	N/A	29.25	48.56	36.73	24.14	22.92	30.74	31.17	N/A	**70.11**
Task macro	14	32.20	N/A	41.99	50.42	48.89	36.94	36.68	42.75	43.91	N/A	**74.68**
AR	100	4.37	9.37	29.44	41.23	40.32	14.52	10.12	31.27	15.20	49.92	**77.30**
TTL	90	42.50	36.88	44.38	31.88	48.75	36.88	43.75	48.75	**49.38**	N/A	43.75
LRU	30	12.17	18.02	21.18	61.12	38.84	26.17	18.80	20.93	44.09	52.61	**73.37**
CR	100	22.00	N/A	22.00	60.00	19.00	19.00	19.00	22.00	16.00	18.00	**86.00**

Cognitive judging prompt, and MAB retains native task metrics and the prescribed capability aggregation (Appendix A.3, Tables 9 and 10). Scores are on a 0–100 scale, separately by benchmark. Query-time generation uses MiMo v2.5 with temperature 0 and thinking disabled unless otherwise specified; main LoCoMo-Plus and LLM-scored MAB tasks use DeepSeek-v4-flash judging. Component comparisons share a scorer; LoCoMo intervals use paired history-cluster bootstrap. Aggregates and differences are computed on unrounded per-item scores and then rounded for display, so a reported difference can deviate by up to 0.01 from the difference of two displayed values. Judge sensitivity appears in Appendix A.5.

4.2 Overall Performance

Finding 1: Causeway improves memory use, not merely memory access. Complete per-category scores, all 14 native MAB task scores, and coverage sensitivity are in Appendix A.3, Tables 7–11. Causeway achieves the strongest overall result on all three suites: 55.77 on LoCoMo, 87.78 on LoCoMo-Plus, and 70.11 on MAB. The LoCoMo-Plus result is especially revealing. Full Context exposes the complete conversation history but reaches only 34.41, whereas Causeway reaches 87.78, and flat lexical and dense retrieval remain below 9 points. The bottleneck is therefore not whether the information exists in the input, but whether the system can identify which historical evidence still constrains the current answer. The same distinction appears on MAB, where Causeway improves over LightMem from 48.56 to 70.11 overall, with particularly large gains on Accurate Retrieval (77.30 vs. 41.23) and Conflict Resolution (86.00 vs. 60.00). Together, these results point to a common regime in which long-term memory becomes difficult: the relevant evidence is present somewhere in history, but must be localized and interpreted under an evolving state.

MAB TTL tests a different use of memory: Causeway averages 82.50 across the five classification tasks but scores 5.00 on ReDial, and because the benchmark gives each group half the TTL weight the result is 43.75. Recommendation requires recovering a set of suitable items rather than a

Table 2: LoCoMo component comparison on the 300 frozen ablation items. Blue superscripts show Overall changes from Full; best scores are bold. The matched protocol is given in Appendix A.1; the corresponding LoCoMo-Plus and MAB conflict controls are reported in Tables 13 and 15.

Method	Overall	Ans. F1	MH	Temp.	Open	SH	Adv.	Ans.Cov.
Full CAUSEWAY	**58.43**	**56.35**	**43.70**	61.47	**39.82**	62.42	**65.67**	95.28
Causal route → BM25	57.19 -1.24	54.74	37.93	62.26	38.18	61.38	**65.67**	95.28
w/o temporal modeling	48.57 -9.86	55.24	41.24	55.48	35.89	**64.42**	25.37	94.85
w/o counterfactual verification	47.87 -10.56	50.05	35.80	**63.64**	30.93	53.80	40.30	**96.57**

single supported fact, and InfBench summarization likewise rewards broad coverage (51.75 versus LightMem's 57.24).

4.3 WHERE DOES CAUSAL MEMORY HELP?

Finding 2: the gains concentrate in evidence localization and evolving-state reasoning. Table 12 aligns benchmark slices that independently probe related memory capabilities, each retaining its native metric against a fixed memory-system reference; native-task contributions are visualized in Figure 3 and length-conditioned conflict results in Appendix A.4, Figures 4 and 5. CAUSEWAY improves MAB AR and CR over LightMem by 36.07 and 26.00 points, respectively, including a 54.00-point gain on multi-hop FactConsolidation. It also leads ApexMem on all four LoCoMo-Plus Cognitive types, including Goal and Value constraints that are not explicit causal questions. LoCoMo Multi-hop separates evidence completeness from state tracking: CAUSEWAY scores 37.39 versus LightMem's 42.38, whereas it leads on updated entity chains in MAB FC-MH; an item-level audit attributes most of that gap to four abstentions (Appendix A.9). Saved traces expose a specific bottleneck: complementary facts can reach the candidate pool but be compressed into a single-fact proof, yielding a supported yet incomplete set answer, so locating a valid dependency need not recover every item required for counting or aggregation (Appendix A.9).

4.4 MECHANISM ABLATIONS

Finding 3: the components exhibit distinct functional signatures. We examine retrieval, temporal modeling, and verification through matched component comparisons (Table 2; conflict-resolution checkpoint lineage in Appendix A.6). Removing temporal modeling or verification reduces matched LoCoMo Overall by 9.86 and 10.56 points, with paired 95% history-cluster intervals of $[-17.21, -2.87]$ and $[-14.07, -7.02]$, whereas replacing the causal route with BM25 changes Overall by -1.24 points (95% CI $[-3.06, +0.83]$) and Answerable F1 by -1.60 ($[-3.00, -0.12]$). Route-level controls on the other two suites show larger changes (Tables 13 and 15): replacing Plus's hierarchical session selector with event-level hashing retrieval lowers Overall from 88.00 to 5.00 and Causal from 94.00 to 8.00. This is a whole-route comparison, in which candidate granularity and routing guidance change and evidence-token lengths are unmatched. On MAB, the BM25 route reduces FC-MH from 76.00 to 16.00 while FC-SH remains intact. A separate fixed-candidate study varies only the ranking formula under a common reader and evidence-budget ceiling (Appendix A.8.1). Removing CE lowers Answerable F1 by 1.87 and Multi-hop F1 by 5.08 points. The 1.16-point answerable advantage over a cosine replacement has a 95% interval crossing zero; removing the negative-control term leaves all answers and scores unchanged. These results support a conditional CE contribution, not superiority over every semantic ranker or an observed negative-control benefit. They do not isolate the contribution of graph-selected background.

Temporal modeling has a different signature. Removing it lowers MAB CR from 84.00 to 34.00, with both FC-MH (76→20) and FC-SH (92→48) declining. About 9.00 of the 9.86 LoCoMo Overall points come from adversarial scoring; Temporal F1 improves by 5.99. On Plus, removing time raises State from 86 to 96 while lowering Overall from 88 to 87, so temporal gains are task-dependent. Verification strongly improves LoCoMo, but on the MAB conflict probe removing the gate recovers one FC-SH answer (24/25 versus 23/25) with FC-MH unchanged, so direct

Table 3: Accuracy versus query-stage LLM usage on the 401-item LoCoMo-Plus Cognitive set. Tokens are thousands per query; "vs. Full" is total tokens relative to Full Context; attempts include failed requests. Construction and judging are excluded; all baselines and input/output splits appear in Table 24. Best values are bold; CAUSEWAY is shaded.

Method	Acc. (%)↑	Total (k tok)↓	vs. Full (%)↓	API attempts / query↓
Full Context	34.41	38.974	100.0	**1.00**
Dense RAG	8.48	4.753	12.2	**1.00**
LightMem	57.61	**1.170**	**3.0**	1.56
MAGMA	17.71	2.562	6.6	1.04
ApexMem*	82.29	6.093	15.6	7.04
CAUSEWAY	**87.78**	8.967	23.0	8.31

current-slot lookup benefits less from an additional rejection step than chain recovery. The components contribute differently to dependency selection, version validity, and answer checking.

4.5 VERIFICATION, ANSWER QUALITY, AND ABSTENTION

Finding 4: verification improves answer quality with a modest coverage change. Appendix A.7, Table 16 reports matched answer quality and response coverage. Verification increases Overall by 10.56 points (paired 95% CI $[7.02, 14.07]$) and Answerable F1 by 6.30 points, while answerable coverage decreases by 1.29 points. Because Answerable F1 counts abstentions as zero, this gain is not an artifact of conditioning on answered questions. The largest improvement is in adversarial refusal performance ($+25.37$ points), while Temporal F1 decreases by 2.17 points. In a separate study on frozen pre-audit states, real-deletion diagnosis versus ordinary evidence checking raises Answerable F1 by 5.16 points, but lowers Overall by 9.32 and adversarial score by 59.70 (Appendix A.8.2). This is a strategy trade-off, not a decomposition of the 6.30-point gain; refusal wording also affects the official score (Appendix A.8.3).

4.6 QUALITY–COST TRADE-OFF

Finding 5: targeted evidence use beats long-context reading at a fraction of the query tokens. Table 3 pairs main-evaluation accuracy with measured query-stage LLM usage (full profile in Appendix A.11.1; LightMem construction costs in Appendix A.11.2). CAUSEWAY gains 53.37 points over Full Context while using 76.99% fewer query-stage tokens. Against ApexMem, it reduces errors from 71 to 49 of 401 (30.99% relative) at $1.47\times$ the tokens and $1.18\times$ the API attempts; against LightMem, it gains 30.17 points at $7.67\times$ the tokens. CAUSEWAY is thus a higher-accuracy, higher-cost operating point among memory systems, yet remains far cheaper than long-context reading. Because these profiles exclude memory construction, they do not establish end-to-end cost parity.

5 CONCLUSION AND OUTLOOK

CAUSEWAY organizes long-horizon memory around the evidence an answer depends on: its bitemporal ledger preserves changing facts and their sources, query-specific evidence selection connects historical dependencies to current questions, and verification traces unsupported answers back to the original record. Across three benchmarks, this combination achieves the strongest overall performance among evaluated methods, with the largest gains in locating evidence, maintaining current states, and recovering persistent constraints. Component studies isolate the contributions of temporal modeling and verification, and the cost analysis shows that targeted reading can outperform full-history prompting with far fewer query tokens. Several directions remain open. First, calibrating the rule-based abstention and replay triggers with conformal or risk-control methods could yield coverage guarantees. Second, caching sensitivity estimates or learning evidence scorers would amortize the per-query cost of answer-dependence audits. Third, a ledger that retains every version raises privacy and deletion questions that bitemporal versioning exposes but does not resolve. Finally, extending source-linked, version-aware memory to tool traces, multimodal observations, and multi-agent stores would test evidence dependence as a general principle for reliable long-horizon agents, judged by influence rather than resemblance.

AI USE STATEMENT

LLM APIs were used as part of the method and experiments: MiMo v2.5 for CAUSEWAY's training-free memory operations and generation, and DeepSeek-v4-flash for specified LLM-based judging. Generative AI tools were also used for language polishing of author-written text. All results and conclusions were reviewed and approved by the authors.

ETHICS STATEMENT

The authors have read and adhere to the ICLR Code of Ethics. The experiments use public benchmark datasets, specified in Appendix A.1, and involve no new human-subject data or personally identifiable information. Because non-destructive memory preserves earlier states, deployments of similar systems should include appropriate consent, access control, retention, redaction, and deletion mechanisms.

REPRODUCIBILITY STATEMENT

An anonymous code repository is linked in the abstract. The memory representation, retrieval, and verification routes are specified in Sec. 3. Question sets, scoring configurations, generation and judging settings, baseline implementation, metrics, and aggregation rules are given in Appendix A.1, Table 4; complete per-category and native-task scores, task weights, and coverage analyses are in Appendix A.3; component studies and their matched protocols are in Appendices A.5, A.6, and A.8; query-stage resource accounting is in Appendix A.11.

REFERENCES

James F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26 (11):832–843, 1983.

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, volume 2024, pp. 55198–55218, 2024.

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*, volume 2024, pp. 9112–9141, 2024.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, 2024.

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, 2026.

Yanyu Chen, Jiyue Jiang, Jiahong Liu, Yifei Zhang, Xiao Guo, and Irwin King. TRACE: Trajectory-aware comprehensive evaluation for deep research agents. In *Proceedings of the ACM Web Conference 2026*, pp. 2524–2534, 2026a.

Yanyu Chen, Jiyue Jiang, Dianzhi Yu, Zheng Wu, Jiahong Liu, Jiaming Han, Xiao Guo, Jinhu Qi, Yu Li, Yifei Zhang, et al. LC-ERD: Mining latent logic for self-evolving reasoning via consistency-regulated reward decomposition. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 496–507, 2026b.

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

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, 2024.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. PerLTQA: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pp. 152–164, 2024.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. LightMem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, volume 2026, pp. 98706–98729, 2026.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37:59532–59569, 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

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

Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. MAGMA: A multi-graph based agentic memory architecture for AI agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36848–36865, 2026.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

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

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

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

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024.

John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pp. 431–450. Elsevier, 1981.

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.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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*, pp. 1–22, 2023.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Vishal Raman, Vijai Aravindh R, and Abhijith Ragav. REMI: A novel causal schema memory architecture for personalized lifestyle recommendation agents. *arXiv preprint arXiv:2509.06269*, 2025.

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.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Richard Snodgrass and Ilsoo Ahn. A taxonomy of time databases. *ACM Sigmod Record*, 14(4): 236–246, 1985.

Saksham Sahai Srivastava. Causal intervention-based memory selection for long-horizon LLM agents. *arXiv preprint arXiv:2605.17641*, 2026.

Ziming Wang. TOKI: A bitemporal operator algebra for contradiction resolution in LLM-agent persistent memory. *arXiv preprint arXiv:2606.06240*, 2026.

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

Yating Wu, Yuhao Zhang, Sayan Ghosh, Sourya Basu, Anoop Deoras, Jun Huan, and Gaurav Gupta. ContextWeaver: Selective and dependency-structured memory construction for LLM agents. *arXiv preprint arXiv:2604.23069*, 2026.

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

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, et al. Mitigating LLM hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*, 2024.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

Chang Yang, Chuang Zhou, Yilin Xiao, Su Dong, Luyao Zhuang, Yujing Zhang, Zhu Wang, Zijin Hong, Zheng Yuan, Zhishang Xiang, et al. Graph-based agent memory: Taxonomy, techniques, and applications. *arXiv preprint arXiv:2602.05665*, 2026.

Xiaohui Zhang, Zequn Sun, Chengyuan Yang, Yaqin Jin, Yazhong Zhang, and Wei Hu. ActMem: Bridging the gap between memory retrieval and reasoning in LLM agents. *arXiv preprint arXiv:2603.00026*, 2026.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. ∞ Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024.

Jingqi Zhou, Sheng Wang, Jingwei Dong, Kai Liu, Lei Li, Jiahui Gao, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. ProReason: Multi-modal proactive reasoning with decoupled eyesight and wisdom. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31650–31679, 2025.

A Appendix

A.1 Evaluation Protocol

Evaluation sets. The main comparison uses fixed, benchmark-specific question sets shared across methods. LoCoMo uses 600 questions selected with seed 42 by stable hashing within each official category under frozen quotas, approximately 30% of the released QA set, spanning ten conversation histories; the category counts are 85 multi-hop, 95 temporal, 50 open-domain, 235 single-hop, and 135 adversarial questions. The selected unit is a question, not a truncated conversation. LoCoMo-Plus uses all 401 Cognitive questions, with 101 causal and 100 each of state, goal, and value questions. MAB uses a seed-42, compute-aware stratified set of 320 QA pairs spanning all 14 native tasks and 54 distinct contexts: 100 AR, 90 TTL, 30 LRU, and 100 CR pairs. Its native-task denominators and context counts are listed in Tables 9 and 10. Selection manifests specify the IDs before the corresponding comparisons; every method uses the same planned IDs, with missing outputs identified separately.

Table 4: Question sets, scoring configurations, and comparison units.

Evaluation	Questions	Scoring / organization
Main comparison		
LoCoMo	600	Token F1 on answerable questions; released adversarial refusal scorer; ten histories.
LoCoMo-Plus Cognitive	401	Binary Cognitive judgment with DeepSeek-v4-flash; complete Cognitive set.
MemoryAgentBench	320	Native task scorers; DeepSeek-v4-flash for LLM-judged tasks; four capability groups.
Component panels and supplementary diagnostics		
LoCoMo v3 ablations	300	Same frozen items as Table 2; paired history-cluster intervals.
Cognitive controls	200	Same-ID predictions scored with `deepseek-flash`, temperature 0, strict binary JSON; Table 13.
MAB CR probes	50	FC-MH and FC-SH, 25 each; Table 15.
Historical Cognitive rejudge	200	Same saved answers rejudged with MiMo; Appendix A.5.
Fixed-candidate ranking	300	Six ranking keys; common candidate pool, top-k, evidence budget, and reader; Appendix A.8.1.
Verification strategy isolation	300	Shared pre-audit drafts and proofs; prompted checking versus executed deletion; Appendix A.8.2.
Earlier one-call diagnostic	40	Eight items per LoCoMo category; seed 42; Appendix A.7.

Generation and judgment. The recorded query-time evaluation uses MiMo v2.5 (model identifier `mimo-v2.5`) at temperature zero with thinking disabled, unless a supplementary experiment specifies otherwise. Main Cognitive judgment and the LLM-scored MAB tasks retain the frozen DeepSeek-v4-flash configuration. The completed Cognitive component rejudgment uses the separately recorded `deepseek-flash` identifier, temperature zero, and a strict parser that accepts only a JSON object with a binary `label` and a `reason`. The historical MiMo rejudgment is a judge-sensitivity comparison on the same saved answers. LoCoMo scoring is deterministic and makes no judge-model call.

Evaluated routes and version control. LoCoMo uses the final v3 reader policy on cached source-grounded graphs. This policy composes the graph answer with a factual quote fallback for public question categories 1, 2, and 4 and a category-blind polar renderer. The policy was revised after development feedback and frozen before full-set composition and scoring. The LoCoMo ablation rows apply the same downstream v3 policy to the corresponding component outputs on 300 frozen IDs: 43 multi-hop, 48 temporal, 25 open-domain, 117 single-hop, and 67 adversarial. The Full row is projected from the final 600-item predictions.

The Cognitive route reuses an immutable LoCoMo base graph and adds a deterministic source-grounded per-question overlay. Its hierarchical selector exposes session synopses under opaque IDs and passes one selected source document to the Cognitive reader. The flat control ranks event-level units with the recorded local hashing-similarity backend, while the temporal control projects the query graph to an atemporal view. These routes use the same questions and source histories; the flat control changes candidate granularity and routing guidance as well as retrieval. Its final top-1 document count does not enforce equal evidence-token lengths.

MAB uses native-task adapters. FactConsolidation maintains current-slot facts and entity-chain retrieval with a deterministic critical-node audit; generic QA/classification, recommendation, and summarization use their corresponding evidence readers. The final 320-item result contains 100 re-run predictions from the revised classification and DetectiveQA routes and 220 retained checkpoints. The CR component comparison covers 25 FC-MH and 25 FC-SH items whose parent checkpoints match the final main run. The earlier 160-item campaign is documented separately in Appendix A.6.

Observed activation and final-output scope. The retained 300 LoCoMo parent traces all contain CE scores and deterministic proof checks; 103 execute node deletion and 229 execute source replay. Common-cause backgrounds are nonempty on 10/300 questions (45/1,795 candidates), and negative-control probes on 6/300 (34/1,795). These are recorded activations, not estimates for the remaining 300 questions. On the complete 600-item final export, postprocessing changes 38 answers: 20 use factual quote fallback and 22 use the polar renderer (routes can overlap). In the detailed subset, 21 outputs undergo postprocessing after core auditing, with no recorded counterfactual re-audit; core-audit status therefore does not certify every emitted answer.

The current Cognitive route uses hierarchical session selection rather than the unified CE scorer/auditor; historical per-module call counts are unavailable. MAB has 200 generic, 100 conflict, ten summarization, and ten recommendation checkpoints. All 100 conflict items record deterministic auditing, with 95 identifying and deleting support; other routes have no recorded CE/deletion audit. These distinctions separate system-level utility from attribution to a particular mechanism.

Input permissions. Public question/task categories guide LoCoMo fallback eligibility and some Plus/MAB compilation and routing. The Plus overlay uses the public cue dialogue; the selector hides cue flags and official cue IDs, while relation/type metadata reaches the compiler. Saved protocols exclude reference answers and judge feedback from generation. Historical Plus exports do not preserve every runtime prompt, so this code-level contract is not a reconstructed per-item prompt audit. Baselines have their own task adapters, not a demonstrated identical metadata-use policy; the comparison is between evaluated systems, not an adapter-free test of graph structure.

Baseline implementation. ApexMem* is a paper-guided reimplementation with an append-only typed SQLite property graph, source and validity annotations, and the four read tools `SchemaViewer`, `EntityLookup`, `GraphSQL`, and `Search`. Prompts, ontology details, and embedding choices are fixed by the local implementation and recorded in its run configuration. Runs use MiMo v2.5 and the frozen benchmark-specific question sets.

MAGMA implementation. MAGMA is the strongest LoCoMo Overall baseline (48.69), a 7.08-point gap from CAUSEWAY. MAGMA Jiang et al. (2026) uses the official implementation (revision `467cb70b67ac`), MiMo v2.5, and local `all-MiniLM-L6-v2` embeddings. The cached vector-index loader was corrected to read `vectors/index.faiss`. The 600 LoCoMo predictions use question text and cached history graphs, with category and reference-answer arguments unset, fixed top-$k = 15$, and one answer candidate. The 135 adversarial and 465 answerable outputs remain in separate source directories and are joined only by their disjoint item IDs for scoring. At scoring time, the exact native refusal `Information not found` is rendered as `No information available.` for adversarial items; the other four types retain their original text.

The released scorer is unchanged. Two answerable refusals receive nonzero lexical overlap, hence the denominator-based definition of Answerable F1 rather than an extra zeroing rule.

Plus reuses LoCoMo graphs by conversation ID; the current cue remains query-local. MAB retains 28 earlier valid answers and 292 resumed answers, with item identity and source hashes checked. Cognitive and native MAB scoring use the same evaluation scripts and aggregation as the other baselines. MAGMA judging on 25 September 2026 requests `deepseek-v4-flash` through the official API and records the returned identifier `deepseek-flash`. For Cognitive, 398 final judgments use a 512-token ceiling; three empty, length-limited responses recover at 2,048 tokens with the same prompt. Only invalid responses are retried, leaving no missing or unparseable final scores. Judge metadata and source hashes accompany the item-level supplement.

Table conventions. All performance scores use a 0–100 scale. In the main table, n counts questions except on macro rows, where it counts categories or tasks. ApexMem* denotes the paper-guided reimplementation. Table 2 and Table 16 use the same 300 LoCoMo component items; the 200-item Cognitive and 50-item MAB CR controls appear in Tables 13 and 15. Table 18 and the verification-strategy study of Appendix A.8.2 reuse these 300 IDs as separate controlled probes, whose readouts differ from the complete-system v3 pipeline. Flat controls use BM25 for LoCoMo/MAB and event-level hashing similarity for Cognitive. Aggregates and differences are computed on unrounded per-item scores and then rounded for display, so a reported difference can deviate by up to 0.01 from the difference of two displayed values.

Metrics and aggregation. LoCoMo Overall is the question-weighted mean of the released per-item scores; Type macro is the unweighted mean over its five categories. Answerable F1 includes every question in the four answerable types, with abstentions retained under the released scorer. Answerable response coverage is the fraction of these questions with a nonempty, non-abstaining response. Cognitive accuracy uses the fixed question denominator; its Type macro gives equal weight to the four types. MAB first averages the native tasks within the prescribed group structure and then gives equal weight to AR, TTL, LRU, and CR. Within TTL, the five many-class classification tasks together receive half of the capability weight, and ReDial receives the other half. MAB Task macro instead averages the 14 native task scores. Table 10 lists the corresponding Overall weights. Aggregates requiring missing task scores are left unavailable.

Comparison units. The LoCoMo intervals resample ten conversation histories with replacement, pool the paired item differences within the sampled histories, and use the 2.5th and 97.5th percentiles of 20,000 draws (seed 42). This preserves dependence among questions from the same history. Coverage sensitivity, native-task context counts, and CR component lineage are documented in Appendices A.3 and A.6.

A.2 Graph-Path Implementation Contract

This section specifies the reference graph path and distinguishes it from the task routes above. The ledger stores source-linked causal claims, temporal relations, and version links; it does not supply structural equations for historical events. Its causal structure is operationally useful for retrieval and inspection without being an identified natural-world SCM.

Scoring and answer labels. The joint prompt receives the question, compiled outcome description and answer format, and two JSON objects, `world_present` and `world_absent`, each containing events and induced edges. The latter removes the candidate events and their incident edges; selected background events remain in both. The implementation retains the candidate-pool union plus background, so disabling background selection need not remove a common cause already in the pool. Outcome proxies are query-matched historical events (up to five positive-relevance targets in the fallback matcher), not newly inferred causal parents of a linguistic answer variable. The implemented common-parent rule also excludes represented descendants, but does not test all paths required for general backdoor identification.

The graph reader asks the model to propose shared canonical answer strings with a probability column for each world. It retains up to six proposed rows; lowercases each string, extracts ASCII alphanumeric tokens, joins them with spaces, and truncates the label to 240 characters. Duplicate labels have their nonnegative masses summed, and each column is normalized separately. A zero-mass world is assigned to `insufficient evidence`. There is no external semantic-equivalence

model: paraphrase alignment depends on the joint prompt. All 3,590 saved world states in the 300-item ranking audit are marked `model_reported_probability`; they are not logits or independent sample frequencies. Seeing both contexts may influence each reported distribution, and independent-world measurement validity has not been established.

The actual joint-prompt contract includes the following instructions and fields (event payloads omitted):

> Compare both intervention worlds using one shared canonical answer space. Use only the supplied source-grounded events and edges. Include insufficient evidence as an answer when appropriate. Probabilities in each world must separately sum to one. Return JSON only.
>
> `answers`: rows of `text`, `present_probability`, and `absent_probability`; plus `presence_support_event_ids` and `absence_support_event_ids`.

The implementation computes TV over the resulting shared labels.

Admission, time, and proof controls. Reference graph-path defaults are $\alpha = (0.34, 0.22, 0.12, 0.14, 0.18)$, edge-admission/proof-edge confidence floor 0.25, candidate score floor 0.12, cost weight $\lambda = 0.04$, conflict weight $\mu = 0.35$, and at most two unrelated probes. These are configured heuristic weights, not fitted or calibrated causal probabilities; budget overrides and native-task policies belong to the respective run configurations. The edge miner caps model-reported temporal, mechanism, and evidence confidence by deterministic gate values and multiplies them. Temporal order uses resolved dates; explicit same-source causal clauses may establish finer precedence within a shared date, whereas mere same-date co-occurrence cannot. Unknown/relative dates use source-time grounding and retain their labels; they are not independently observed event times.

The edge-mining prompt requests an exact source quote for `explicit_causes`, `explicit_enables`, or `explicit_prevents`, and assigns plausible, temporal-only, similarity-only, or commonsense causation to `causal_hypothesis`. It returns at most 12 proposed edges per call through `emit_edge_batch`, including endpoints, relation, polarity, mechanism, source quote, and the three gate confidences. An explicit injury-to-trampoline-replacement clause can therefore propose an admitted causal claim; two merely adjacent statements sharing an entity remain hypotheses. These illustrate the rule, not independently annotated graph accuracy. Explicit prevention carries negative polarity; admitted relations remain defeasible claims rather than certified facts.

Proof construction searches candidate combinations under score, cost, and conflict constraints. On branches requiring counterfactual checks, the selected proof is evaluated jointly; a non-counterfactual branch can use the maximum selected candidate CE as its proof effect. Thus a high-scoring singleton can saturate the objective while complementary facts add cost (Appendix A.9). Temporal admissibility is query-specific: causal explanations can retain earlier occurrences even after their state interval ends. Source coverage of emitted claims and coverage of all requested answer elements are distinct; the former does not imply set/count completeness.

Audit contract. Applicable removal tests exclude the target node and incident edges from the reader-visible proof. A polarity test retains the endpoints, negates polarity, maps the relation to its configured opposite (e.g., causes to prevents), and replaces the mechanism with an explicit counterfactual-flip marker. It is a graph-evidence perturbation, not a coherently regenerated historical world. A single-node test can miss redundant or jointly necessary support. Source validity, contradiction checks, and replay complement answer-dependence tests; neither an answer change nor an unchanged answer alone verifies truth. Historical-belief reconstruction is enabled by the two stored time axes, but the reported no-temporal controls do not isolate valid time from transaction time.

A.3 Complete Scores, Coverage, and Aggregation

This appendix provides complete scores, task weights, coverage analyses, and supplementary diagnostics. Evaluation sets and version lineage are specified in Appendix A.1; the accompanying audit supplies unrounded aggregates and source hashes.

Answerable quality and uncertainty. On all 465 answerable items of the main 600-question LoCoMo set, CAUSEWAY scores 53.69 F1 versus LightMem's 48.25; these are computed from unrounded item scores, not from the 300-item ablation. Table 5 also reports lexical response coverage. Coverage uses the frozen audit's explicit-refusal phrases and is a reproducible text diagnostic, not an independent semantic refusal annotation. These main-comparison audit intervals use 5,000 paired

Table 5: Main LoCoMo answerable results, $n = 465$.

Method	Answerable F1	Response coverage (%)
A-Mem	14.89	30.75
Mem0	16.04	27.96
LightMem	48.25	98.49
MAGMA	38.74	72.69
BM25	35.08	71.18
Dense RAG	15.13	41.08
Hybrid RAG	34.88	70.32
Truncated Context	13.87	38.49
Full Context	22.63	54.84
ApexMem*	22.41	96.99
CAUSEWAY	53.69	94.84

Table 6: Exploratory paired Overall differences, CAUSEWAY minus baseline.

Benchmark	Baseline	n	Difference	95% CI
LoCoMo	LightMem	600	+8.05	[+3.16, +13.03]
LoCoMo	ApexMem*	600	+37.57	[+34.02, +41.26]
LoCoMo	MAGMA	600	+7.08	[+2.43, +11.52]
LoCoMo-Plus	LightMem	401	+30.17	[+26.12, +33.66]
LoCoMo-Plus	ApexMem*	401	+5.49	[-0.27, +11.19]
LoCoMo-Plus	MAGMA	401	+70.07	[+64.49, +74.69]
MAB	LightMem	320	+21.55	[+14.62, +25.86]
MAB	ApexMem*	320	+45.97	[+39.22, +50.48]
MAB	MAGMA	320	+33.38	[+27.55, +38.64]

cluster-bootstrap draws (seed 42), distinct from the 20,000-draw component intervals: ten histories for LoCoMo/Plus and 54 MAB contexts, resampled within native tasks with the prescribed hierarchy recomputed. Tasks with one context contribute no estimable within-task cluster variance. The intervals condition on the frozen subsets and judging snapshots, with no multiplicity correction; they do not establish full-benchmark or cross-judge generalization. In particular, the Plus advantage over ApexMem has an interval crossing zero.

Table 7: Complete LoCoMo category scores, using the same predictions and scoring as Table 1.

Type / aggregate	**Trunc. ctx.**	**Full ctx.**	**BM25**	**Dense**	**Hybrid**	**A-Mem**	**Mem0**	**Light Mem**	**MAGMA**	**Apex***	**Causeway**
LoCoMo											
Multi-hop	18.20	21.92	22.50	11.05	21.65	11.50	13.20	**42.38**	23.22	23.45	37.39
Temporal	7.28	25.94	22.39	7.30	21.07	4.84	16.76	54.34	35.51	22.79	**57.38**
Open-domain	6.45	8.79	8.07	11.65	6.36	4.96	4.67	27.78	17.74	10.60	**34.19**
Single-hop	16.55	24.49	50.51	20.52	51.32	22.30	19.19	52.26	50.13	24.40	**62.24**
Adversarial	73.33	74.81	62.22	81.48	61.48	89.63	**97.78**	45.93	82.96	3.70	62.96

Apex* denotes the paper-guided reimplementation.

Table 12 re-aligns the scores above into cross-benchmark capability slices; it is the reference for Finding 2 in Section 4.3 and introduces no new predictions.

Table 8: LoCoMo-Plus Overall and category scores, using the same predictions and judging snapshot as Table 1.

Type / aggregate	Trunc. ctx.	Full ctx.	BM25	Dense	Hybrid	A-Mem	Mem0	Light Mem	MAGMA	Apex*	Causeway
LoCoMo-Plus Cognitive											
Overall	33.92	34.41	7.23	8.48	6.98	14.71	8.73	57.61	17.71	82.29	**87.78**
Causal	42.57	43.56	11.88	7.92	8.91	20.79	15.84	58.42	31.68	87.13	**94.06**
State	35.00	33.00	6.00	11.00	8.00	6.00	8.00	67.00	15.00	84.00	**86.00**
Goal	32.00	34.00	3.00	3.00	3.00	15.00	5.00	50.00	12.00	79.00	**82.00**
Value	26.00	27.00	8.00	12.00	8.00	17.00	6.00	55.00	12.00	79.00	**89.00**

Apex* denotes the reimplementation. Two CAUSEWAY judge responses require parser review (Appendix A.5).

Table 9: All 14 native MAB tasks. n is the planned QA-pair count. Mem0 FC-MH/SH each cover 46/50 items; Full Context has no ReDial score.

Native task	n	Trunc. ctx.	Full ctx.	BM25	Dense	Hybrid	A-Mem	Mem0	Light Mem	MAGMA	Apex*	Causeway
Accurate Retrieval												
EventQA	35	28.57	68.57	60.00	17.14	62.86	8.57	8.57	17.14	65.71	11.43	**91.43**
LongMemEval	45	22.22	**71.11**	37.78	13.33	42.22	8.89	8.89	17.78	35.56	6.67	57.78
RULER-QA1	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	80.00	10.00	20.00	**100.00**
RULER-QA2	10	10.00	**60.00**	20.00	10.00	20.00	0.00	20.00	50.00	50.00	20.00	**60.00**
Test-Time Learning												
BANKING77	16	**93.75**	**93.75**	75.00	81.25	75.00	62.50	81.25	50.00	87.50	75.00	75.00
CLINC150	16	**93.75**	**93.75**	**93.75**	87.50	**93.75**	87.50	75.00	87.50	81.25	87.50	87.50
NLU	16	81.25	**100.00**	81.25	81.25	75.00	87.50	81.25	75.00	93.75	87.50	93.75
TREC-Coarse	16	75.00	**81.25**	75.00	**81.25**	62.50	56.25	75.00	50.00	56.25	68.75	75.00
TREC-Fine	16	75.00	**81.25**	43.75	56.25	56.25	56.25	56.25	31.25	68.75	50.00	**81.25**
ReDial	10	15.00		15.00	10.00	**25.00**	15.00	0.00	5.00	20.00	0.00	5.00
Long-Range Understanding												
DetectiveQA	20	50.00	50.00	25.00	20.00	20.00	15.00	25.00	65.00	40.00	50.00	**95.00**
InfBench-Sum	10	38.17	55.22	17.37	17.61	21.86	9.35	11.05	**57.24**	37.68	2.33	51.75
Conflict Resolution												
FC-MH	50	0.00	0.00	0.00	0.00	0.00	0.00	2.17	26.00	0.00	2.00	**80.00**
FC-SH	50	32.00	36.00	44.00	38.00	44.00	44.00	43.48	**94.00**	38.00	36.00	92.00

Metrics, weights, and context counts are given in Table 10. MAB spans 54 distinct contexts.

Table 10: Fixed MAB task weights and their arithmetic contributions to the +21.55-point Overall difference from LightMem. K is the number of source contexts.

Task	n	K	Metric	Overall weight	Gain contrib.
Accurate Retrieval					
EventQA	35	13	SubEM	6.25%	+4.64
LongMemEval	45	5	LME judge	6.25%	+2.50
RULER-QA1	10	1	SubEM	6.25%	+1.25
RULER-QA2	10	1	SubEM	6.25%	+0.62
Test-Time Learning					
BANKING77	16	1	Exact Match	2.50%	+0.62
CLINC150	16	1	Exact Match	2.50%	+0.00
NLU	16	1	Exact Match	2.50%	+0.47
TREC-Coarse	16	1	Exact Match	2.50%	+0.62
TREC-Fine	16	1	Exact Match	2.50%	+1.25
ReDial	10	1	Recall@5	12.50%	+0.00
Long-Range Understanding					
DetectiveQA	20	10	Exact Match	12.50%	+3.75
InfBench-Sum	10	10	Sum. score	12.50%	−0.69
Conflict Resolution					
FC-MH	50	4	SubEM	12.50%	+6.75
FC-SH	50	4	SubEM	12.50%	−0.25

Eight single-context tasks account for 37.5% of Overall; the recorded bootstrap supplies no within-task cluster uncertainty for these tasks. Contributions are computed on unrounded scores and rounded for display, so the displayed column sums to 21.53 rather than the exact total of 21.55.

Table 11: MAB coverage sensitivity: Overall on the 312 questions scored for every listed method, the intersection with Mem0's available questions. Main-set Overall and Task macro are given in Table 1 and preserve this ordering. Full Context is omitted because it covers only 302/312, having no ReDial score, and Mem0 has no complete main-set aggregate, so the intersection is its only comparable score.

Method	Same312 Overall
Trunc.	31.51
BM25	29.19
Dense	23.33
Hybrid	30.67
A-Mem	20.19
Mem0	21.77
LightMem	48.50
MAGMA	36.87
Apex*	23.74
CAUSEWAY	**71.16**

The intersection excludes eight 262K conflict questions, on which CAUSEWAY scores 3/8.

Table 12: Cross-benchmark capability signatures. The reference is fixed per benchmark: LightMem for LoCoMo and MAB, and ApexMem for LoCoMo-Plus. Δ denotes CAUSEWAY minus the corresponding reference in percentage points, computed before rounding the displayed scores.

Capability	**Benchmark slice**	**Causeway**	**Ref.**	Δ
Evidence localization	LoCoMo Single-hop	62.24	52.26	+9.97
	MAB AR	77.30	41.23	+36.07
Evolving-state maintenance	LoCoMo Temporal	57.38	54.34	+3.04
	LoCoMo-Plus State	86.00	84.00	+2.00
	MAB CR	86.00	60.00	+26.00
Dependency reasoning	LoCoMo Multi-hop	37.39	42.38	-4.99
	LoCoMo-Plus Causal	94.06	87.13	+6.93
	MAB FC-MH	80.00	26.00	+54.00
Persistent constraints	LoCoMo-Plus Goal	82.00	79.00	+3.00
	LoCoMo-Plus Value	89.00	79.00	+10.00

Aggregation sensitivity. On the shared 312 IDs, CAUSEWAY scores 71.16 and LightMem 48.50. On the original 320-item set, their Task macro scores are 74.68 and 50.42. Both comparisons preserve the ordering. The higher intersection score reflects removal of the eight difficult 262K items, so the main comparison retains all 320 planned questions.

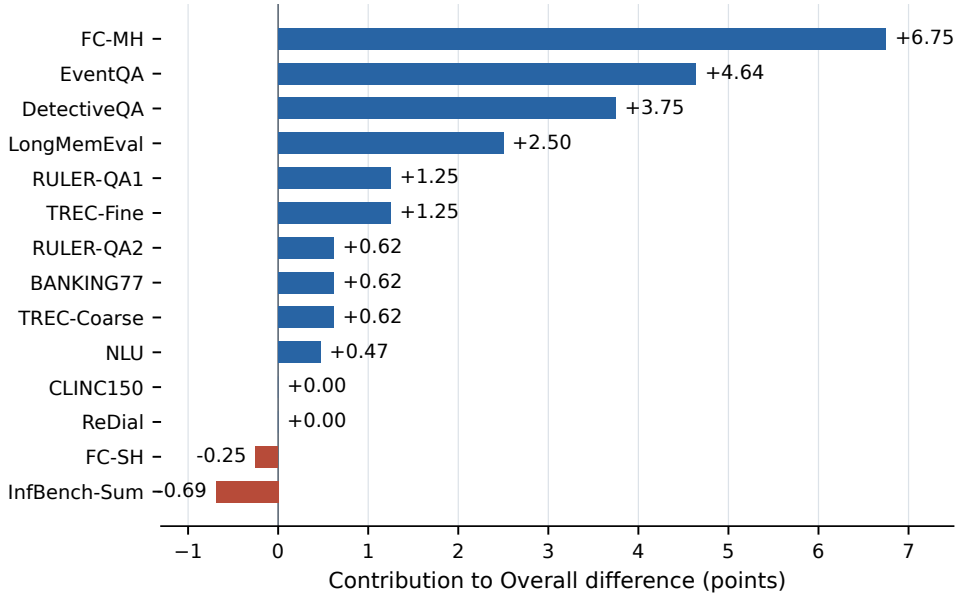

Figure 3: Native-task contributions to the +21.55-point MAB Overall difference from LightMem, using the weights in Table 10.

A.4 Length-Conditioned Conflict Results

The primary CR score averages QA pairs within each native task, then averages SH and MH. The following figures condition on task and context length, with one source context per cell. Length and context identity therefore vary together. The supplementary equal-length macro uses a different weighting from the main CR score.

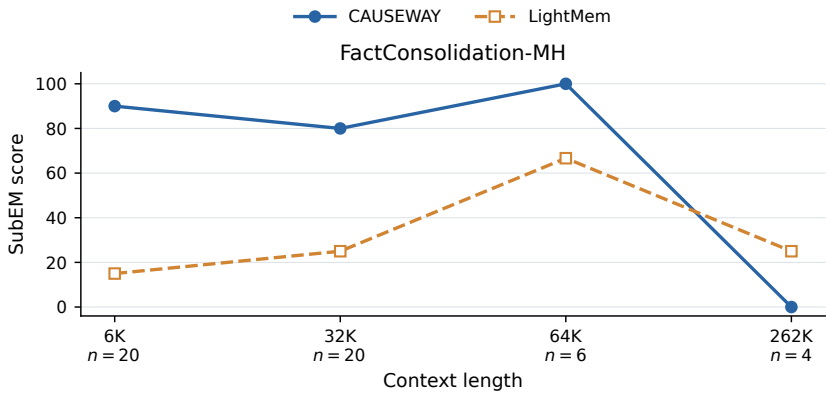

Figure 4: FC-MH scores by recorded context length; each cell corresponds to a different source context.

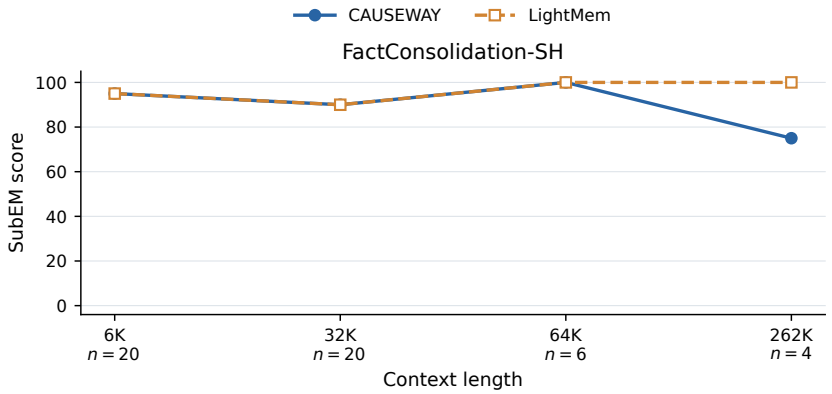

Figure 5: FC-SH scores by recorded context length under the same counting rule as FC-MH.

A.5 Judge Sensitivity and Route Replacement

Table 13 reports the LoCoMo-Plus component comparison discussed in Section 4.4. All three configurations use the same 200 questions and a strict `deepseek-flash` judge at temperature zero; all 600 saved judgments parse as binary JSON. Full uses hierarchical session selection; the flat route uses event-level hashing retrieval. Candidate units, routing guidance, and evidence-token lengths differ (Appendix A.1).

Table 13: Current strict DeepSeek comparison on the frozen Plus subset: 200 items, 50 per type. Exact refusals count the saved `insufficient evidence` output.

Configuration	Overall	Causal	State	Goal	Value	Exact refusals
Full cognitive route	88.00	94.00	86.00	82.00	90.00	14/200
Flat retrieval control	5.00	8.00	4.00	2.00	6.00	120/200
w/o temporal modeling	87.00	88.00	96.00	80.00	84.00	15/200

The same saved Full answers score 176/200 (88.00) with the component judge, 175/200 (87.50) under the original main-table DeepSeek labels, and 162/200 (81.00) under strict MiMo rejudgment. These are judge snapshots of fixed answers. The main 401-item result retains its original 87.78 score.

Table 14: Historical strict MiMo rejudgment on the same 200 Plus items (50 per type). All 400 saved judge responses parse as binary JSON.

Route	**Overall**	**Causal**	**State**	**Goal**	**Value**	**Exact refusals**
Full cognitive route	**81.00**	**84.00**	**82.00**	**72.00**	**86.00**	14/200
Semantic-flat route	1.50	4.00	0.00	0.00	2.00	120/200

Full has 162 correct answers versus 3 for the flat route, with paired wins/ties/losses of 159/41/0. The flat route is correct in 3/80 non-refusals.

Main DeepSeek and strict MiMo judgments agree on 187/200 Full answers (93.5%): 162 both-correct, 25 both-wrong, and 13 DeepSeek-correct/MiMo-wrong. The exploratory history-cluster interval for the MiMo-minus-DeepSeek difference is $[-9.50, -3.90]$ points. This measures judge sensitivity on fixed answers; component effects use a common judging snapshot. This rejudge does not compare the closest memory baselines under a second judge, so it cannot establish judge-robust superiority over ApexMem or LightMem.

The main 401-item CAUSEWAY snapshot contains 399 strict saved judge responses and two requiring parser review. Changing both binary labels would alter accuracy by at most 0.50 points; the

two responses remain unadjudicated. Baseline raw judge responses are incompletely retained, so comparable parser uncertainty is unavailable.

All component conclusions use the common strict DeepSeek snapshot in Table 13; superseded parser-dependent scores are retained in the artifact audit.

A.6 Conflict-Resolution Component Controls

Table 15: MAB CR controls on 50 frozen questions, scored with native SubEM. Full checkpoints are byte-identical to their main-result parents; graph and variant lineage are verified.

Configuration	CR (50)	FC-MH (25)	FC-SH (25)
Full projection	84.00	**76.00**	92.00
No-causal route	58.00	16.00	**100.00**
No-temporal view	34.00	20.00	48.00
No verification gate	**86.00**	**76.00**	96.00

The table covers CR only; the wider campaign's 110 unchanged generic-route predictions are excluded.

The no-causal route uses BM25 top-16 current-slot facts without entity-chain expansion. The no-temporal route uses an all-version static view without latest-slot or time-valid filtering. The no-verification variant submits saved pre-audit reader outputs. The audit verifies parent, graph, and intervention lineage for these 50 CR questions.

The CR subset has matched parent checkpoints. Non-CR routes were subsequently revised, so the earlier generic-route outputs are not used to infer component effects for the final system.

A.7 Verification Behavior and the Earlier One-Call Diagnostic

Table 16 gives the matched LoCoMo verification comparison discussed in Section 4.5; both rows use the 300 frozen component IDs and the downstream v3 reader policy of Appendix A.1. Among the 233 answerable items, 222 are answered in both conditions: their answer changes contribute $+6.73$ points, while three additional abstentions contribute -0.43 points.

Table 16: Verification on LoCoMo. Red superscripts show changes from the unverified system; best scores are bold. The matched protocol is given in Appendix A.1.

Method	Overall	Answerable F1	Adversarial score	Answerable coverage	Temporal F1
w/o verification	47.87	50.05	40.30	**96.57**	**63.64**
Full CAUSEWAY (Ours)	**58.43** $+10.56$	**56.35** $+6.30$	**65.67** $+25.37$	95.28 -1.29	61.47 -2.17

The remainder of this appendix reports an earlier exploratory diagnostic that uses a prompted deletion instruction rather than an independently executed reduced world; Appendix A.8.2 tests actual deletion under shared drafts and proofs.

Table 17: Exploratory 40-question diagnostic with matched initial answers and evidence. All rows use the same category-blind refusal renderer. The provisional row is a saved pre-check output, distinct from the system-level no-verification ablation.

Path	Raw All	Rendered All	Ans. F1	Adv.	Ans. refusals
Saved provisional output	**15.74**	**30.74**	**19.67**	75.00	19/32
General evidence self-check	7.60	27.60	9.50	**100.00**	25/32
Counterfactual self-check	11.89	29.39	14.86	87.50	21/32

Counterfactual minus general on Rendered All: +1.79 points, with a 20,000-draw history-cluster interval of $[-7.76, +11.63]$ and 3/36/1 wins/ties/losses.

The diagnostic uses seed 42 to freeze eight items per LoCoMo category across ten histories. Each arm receives identical initial answers and evidence and uses one successful MiMo v2.5 call at temperature zero with a 256-token output ceiling. General and counterfactual input totals are 123,957 and 125,117 tokens, respectively (0.94% difference). Two gateway 429 attempts are recorded separately from the 80 successful calls; graphs are reused.

Payloads contain event ID, document ID, and text, with 600 characters per excerpt and approximately 6,000 characters of serialized evidence. Source dates are not a dedicated field, and payload sufficiency was not independently labeled. Some refusals may therefore reflect missing evidence rather than verifier conservatism.

The counterfactual arm uses a prompted necessity stress test, not the multi-call removal/replay pipeline. Four pairs have nonzero score differences; excluding history conv-42 changes the mean difference from +1.79 to -2.71 points. Two outputs declare `keep` while replacing a refusal with a factual answer, so action labels and observed text changes are recorded separately.

The released adversarial scorer rejects the raw refusal string `insufficient evidence`. Rendering explicit abstentions as `No information available` changes the scored refusal wording rather than the underlying evidence; all three rows use the same normalization. The provisional output scores 30.74 after rendering versus 15.74 raw. Exploratory paired intervals for general-minus-provisional and counterfactual-minus-provisional are $[-9.60, +4.88]$ and $[-7.89, +5.22]$ points. This diagnostic separates one-call checking of fixed, excerpted evidence from the system-level verification route, which can replay sources and change the final answer. It supports neither replacing that route with a one-call self-check nor ranking the two self-check prompts reliably.

A.8 Controlled Mechanism Studies

The following studies reuse the 300 frozen LoCoMo component IDs but evaluate distinct controlled readouts. The first isolates the ranking formula; the second compares verification strategies from common pre-audit states. Neither replaces the complete-system predictions in the main tables.

A.8.1 Fixed-Candidate Interventional Ranking

Common inputs and scope. The six formulas share 300 frozen LoCoMo questions (43 MH, 48 Temporal, 25 Open, 117 SH, 67 Adversarial) and 1,795 historically measured candidates, before final score-dependent admissibility; unmeasured proposals are excluded equally across arms. The full ranking formula is a control within this readout. The complete-system Full score on the same IDs is 58.43 Overall and 56.35 Answerable F1, obtained by projecting the 600-item final v3 predictions; its verifier, source replay, proof selection, and category fallback are not part of the ranking-only readout. The two controls answer different questions and are not substituted for each other.

Fixed selection and readout. Every formula selects top-$k = 3$; 34 items have only one candidate and use it in all arms. Candidate texts are frozen before ranking at 640 MiMo-tokenizer tokens each, under a 2,048-token evidence ceiling. This matches maximum budgets rather than padding short evidence. The reader is MiMo v2.5 at temperature zero with a 256-token output ceiling. Its prompt receives the question and selected evidence, without arm names, scores, categories, gold answers, or historical predictions. Identical prompts share a generated response: 1,800 arm–item positions yield 854 unique requests. There is no proof reranking, verifier, source replay, or category fallback in this readout.

Ranking definitions. Let $b = 0.22\,\mathrm{PathRel} + 0.12\,\mathrm{TempValid} + 0.14\,\mathrm{conf}$ and $n = 0.18\,\mathrm{Spurious}$. Full uses $b + 0.34\,\mathrm{CE} - n$; removing CE uses $b - n$; Base uses b; cosine and BM25 replace only CE; removing the negative control uses $b + 0.34\,\mathrm{CE}$. Dense cosine uses local `all-MiniLM-L6-v2`, averages embeddings of 224-wordpiece chunks, unit-normalizes the result, and clips negative cosine values to zero. BM25 uses the same candidate text with $k_1 = 1.2$, $b = 0.75$, normalized by the within-query maximum. Replacement weights are fixed before generation.

Selection and output changes. Relative to Full, top-k sets change on 128, 132, 179, 210, and 4 questions for no-CE, Base, cosine, BM25, and no-negative-control, respectively. Final answer text changes on 55, 55, 57, 66, and 0 questions. The corresponding Full wins/alternative wins/ties

Table 18: Complete controlled-ranking scores. Full formula denotes the ranking readout, not the complete-system Full.

Ranking	Overall	Ans. F1	MH	Temp.	Open	SH	Adv.
Full formula	42.84	30.70	21.98	21.33	12.30	41.67	85.07
w/o CE	41.72	28.83	16.90	25.78	8.16	38.88	86.57
Base only	41.72	28.83	16.90	25.78	8.16	38.88	86.57
CE → cosine	42.27	29.53	20.41	21.68	13.58	39.52	86.57
CE → BM25	41.01	27.91	17.37	20.61	6.74	39.31	86.57
w/o negative control	42.84	30.70	21.98	21.33	12.30	41.67	85.07

Table 19: Full-formula minus alternative differences and history-cluster 95% intervals.

Alternative	Ans. F1 difference	95% CI	MH difference	95% CI
w/o CE	$+1.87$	$[0.16, 4.13]$	$+5.08$	$[0.86, 10.29]$
Base only	$+1.87$	$[0.16, 4.13]$	$+5.08$	$[0.86, 10.29]$
CE → cosine	$+1.16$	$[-0.62, 3.69]$	$+1.57$	$[-3.69, 7.82]$
CE → BM25	$+2.78$	$[0.31, 5.58]$	$+4.61$	$[-3.11, 12.70]$
w/o negative control	0.00	$[0.00, 0.00]$	0.00	$[0.00, 0.00]$

are 23/19/258, 23/19/258, 21/20/259, 27/26/247, and 0/0/300. Removing the negative-control term changes four selected sets but no final answer, providing no observed end-task benefit in this sample. Temporal and adversarial point estimates improve under some alternative rankings; Overall intervals include zero. These category responses characterize where CE contributes, while the answerable and MH contrasts quantify its clearest gains.

All 854 unique responses complete: 853 are strict JSON; one hits the 256-token limit after emitting a complete answer string but before finishing its explanation. A common label-blind prefix parser recovers the already emitted answer. Candidate selection and output budgets are unchanged. Intervals resample ten histories 20,000 times with seed 42, using the paired item differences and percentile endpoints. These are exploratory intervals without multiplicity correction.

A.8.2 Matched Multi-Call Verification Strategies

System-level verification decomposition. In the separate complete-system ablation, 222 of 233 answerable questions receive answers both with and without verification; their answer changes contribute $+6.7275$ points. Eight questions are refused in both conditions, three become refusals, and none switch from refusal to answer. The three additional refusals contribute -0.4292 points, yielding $+6.2983$ points overall on the fixed answerable denominator. This decomposition describes the historical complete-system pair; the following matched study compares newly executed diagnostic strategies from common intermediate states.

Snapshot contract. The study uses the same 300 IDs and saved graph-stage `provisional_answer` and `pre_replay_proof` records, before the final v3 replay/fallback composition. All parent checkpoint hashes match their frozen manifest. A preserves the initial answer. B performs five explicit checks: source support, temporal validity, identity, conflicts, and omitted evidence. C independently answers after a real evidence deletion. These are newly implemented matched strategies on historical states, not a replay of every historical audit prompt or a replacement for the main-system predictions.

Deletion selection and visibility. The frozen selector ranks visible proof events by the number of answer claims they support, then computation membership, proof order, and event ID. C removes the first event and its incident proof edges. It receives only the question, the deleted ID, and the remaining event/edge records; it receives no original answer, claim-support summary, computation summary, or replay source documents. Other event records are retained, including duplicate support. The deletion is defined at the event-record level; semantic redundancy across surviving records is not treated as an error. Of the 300 snapshots, 294 contain a deletable event. The remaining six execute the unchanged/empty-world diagnostic and stay in the full denominator.

Matched multi-step execution. B and C each make one diagnostic call (512 output tokens) and one final adjudication call (384 output tokens), using MiMo v2.5 at temperature zero with thinking disabled. The snapshot includes proof kind, support mappings, source quotes, entities, and valid/transaction times; the full serialized proof input is capped at 8,192 tokens, with an observed maximum of 7,534. The saved replay-eligibility flag is fixed across strategies: 229 items call a shared source reader with a 256-token output ceiling. Its saved candidate-document inputs have a maximum of about 10.6k tokens and are withheld from the deletion diagnostic. Both final adjudicators receive the same full proof and shared replay answer/explanation, alongside their own diagnostic. The adjudication template is common, with a strategy tag identifying the diagnostic type.

B emits separate findings for the five checks and a proposed action. C returns an independently solved answer, visible support IDs, an evidence-insufficiency flag, and a reason. Answer change is defined by lowercasing and removing punctuation/extra whitespace; the stored change statistic is diagnostic, not an automatic rejection rule. Both final adjudicators may keep the exact initial answer, correct it, or abstain. An unchanged deletion answer is compatible with redundant support; a changed answer is not itself a correctness label. Consequently, C–B estimates the effect of a diagnostic-and-adjudication strategy, not deletion in isolation from diagnostic format or a gold measure of faithfulness.

Table 20: Complete raw verification scores under the original rendering rules. One shared replay failure contributes zero to both B and C; every planned question is retained.

Strategy	Overall	Ans. F1	MH	Temp.	Open	SH	Adv.	Ans. cov.
A: initial	14.02	18.05	8.16	35.95	1.76	17.82	0.00	36.48
B: ordinary	40.53	32.02	18.50	26.28	7.03	44.68	70.15	57.94
C: deletion	31.21	37.18	23.66	36.53	10.65	48.09	10.45	75.11

Primary comparison and uncertainty. Table 21 reports C–B differences on fixed denominators; abstentions and the shared failure score zero. The designated primary comparison, Answerable F1, improves by 5.16 points, as does Multi-hop F1. Overall and adversarial scores decrease by 9.32 and 59.70 points. Open-domain and Single-hop intervals cross zero. The paired percentile intervals resample ten histories 20,000 times with seed 42 and are exploratory, without multiplicity correction.

Table 21: Real-deletion strategy C minus ordinary verification B. Positive differences favor C; scores and intervals are in percentage points.

Slice	n	C–B	History-cluster 95% CI
Answerable (primary)	233	$+5.16$	$[+1.25, +7.96]$
Overall	300	-9.32	$[-14.60, -4.95]$
Multi-hop	43	$+5.16$	$[+1.31, +9.78]$
Temporal	48	$+10.24$	$[+1.59, +17.67]$
Open-domain	25	$+3.61$	$[-4.74, +15.00]$
Single-hop	117	$+3.41$	$[-4.48, +8.58]$
Adversarial	67	-59.70	$[-68.52, -52.63]$

Answer quality and refusal trade-off. B and C improve Answerable F1 over A by 13.97 and 19.13 points, respectively. Their full-set improved/degraded/unchanged counts are 120/18/162 and 106/9/185. These A-to-checked comparisons include the added replay and readout workflow. C's answerable coverage is 75.11%, versus B's 57.94%. Final keep/correct/abstain actions are 55/107/138 for B and 100/186/14 for C; keeping an initial refusal remains a refusal, so action counts are not coverage. The original rendering rules are retained: an explicit abstain action emits the scorer-recognized refusal, while keep preserves the original text. The released adversarial scorer does not recognize every refusal wording, so this score reflects both refusal behavior and its surface form. C improves answerable quality but does not dominate B across metrics. Combining deletion diagnosis with ordinary support and refusal checks remains a proposed extension, not an evaluated combined system.

Execution and interpretation limits. All 300 ordinary diagnostics, 300 deletion diagnostics, and 600 final adjudications parse successfully. Of 229 shared replay items, 228 succeed; one fails the JSON schema on both its initial response and a fresh retry and contributes zero to both B and C. Each checked arm has 829 logical call positions (300 diagnostics, 229 shared replay positions, and 300 adjudications); physical execution shares replay results and records 1,430 non-cache responses including the retry. Equal call and output ceilings do not imply equal realized token usage. These diagnostic variants reuse historical graphs and initial answers. Their absolute scores do not replace the main-system results, and the C–B gain cannot be treated as a portion of the separate system-level 6.30-point verification gain. Answer-quality improvement is not an independent measurement of faithfulness or real-world causal identification.

A.8.3 Refusal-Form Sensitivity of the Verification Comparison

We retain the original official scores and apply one additional, gold-independent offline rendering rule to all B/C outputs: lowercase, collapse whitespace, strip terminal punctuation, and match the whole string against `insufficient evidence`, `no information available`, `not enough information`, `cannot be determined`, `can't be determined`, `not mentioned`, `not provided`, or `unknown`. Matched outputs become `No information available`; factual and mixed outputs remain unchanged. Recognition receives neither category nor reference answer; references are used only afterward by the unchanged official scorer. This is a post-hoc lexical sensitivity check, not a comprehensive semantic refusal judge. Normalization

Table 22: Original and uniformly rendered B/C scores on the same 300 items.

Strategy	Rendering	Overall	Answerable F1	Adversarial
B: ordinary	original	40.53	32.02	70.15
B: ordinary	normalized	42.87	32.02	80.60
C: deletion	original	31.21	37.18	10.45
C: deletion	normalized	40.54	37.18	52.24

changes 13 B outputs and 78 C outputs; all changed strings are the exact refusal `insufficient evidence`. The C-minus-B Overall difference becomes -2.32 rather than -9.32, and the adversarial difference -28.36 rather than -59.70. Answerable F1 and its $+5.16$ difference are unchanged. Thus refusal form explains a substantial part, but not all, of this scoring gap under the stated rule. The original scores remain the primary reported evaluation; normalization does not demonstrate evidence faithfulness.

A.9 Multi-hop Completeness and Task-Specific Trade-offs

The post-hoc audit replays the unchanged LoCoMo scorer on all 85 Multi-hop items. Table 23 partitions the original scores by explicit abstention status; it neither drops questions nor changes the main metric.

Table 23: Decomposition of the original Multi-hop gap. Contributions use all 85 items as the denominator; columns report F1 on a 0–100 scale.

Response group	n	CAUSEWAY	LightMem	Gap contribution
Neither response explicitly abstains	81	39.24	40.15	-0.87
CAUSEWAY explicitly abstains	4	0.00	87.50	-4.12
All Multi-hop items	85	37.39	42.38	-4.99

Compression can discard complementary facts. For `conv-26:q60`, both clarinet and violin occur in candidates, but the query is compiled as `fact/short`, the selected proof retains only violin, and the supported partial answer passes audit. For `conv-44:q51`, candidates contain several pet-related items, but the proof retains only personalized tags. In the non-counterfactual selection branch, proof effect uses the maximum candidate CE: once one fact attains CE=1, adding complementary facts need not increase that term but can increase cost. These saved traces locate a

completeness bottleneck before verification. A check that every emitted claim has support does not ensure that every requested item was emitted.

Recovery and answer contracts. Three of the four explicit abstentions have no admitted candidates, and source replay misses required historical sessions. The fourth also involves inconsistent entity names between the question and dialogue. Turning off the verifier alone does not repair these upstream omissions: on the matched 43-item MH ablation, removing verification lowers F1 from 43.70 to 35.80. Completeness-aware evidence selection, event deduplication, and explicit set/count contracts address the observed failure modes more directly.

Selective support versus broad coverage. The same distinction helps interpret recommendation and summarization results: selecting a compact supported answer differs from covering a recommendation set or a long document. For FC-SH, a direct current-slot answer can already be sufficient, consistent with the flat control's 25/25 and Full's 23/25; for FC-MH, the corresponding 4/25 versus 19/25 shows the value of retaining the connecting chain. On LoCoMo adversarial questions, high refusal performance alone also does not imply better answerable recall: Mem0 scores 97.78 on adversarial items but remains below CAUSEWAY in each answerable category.

A.10 ILLUSTRATIVE SOURCE-LEVEL CASES AND METRIC LIMITS

Figure 6 illustrates source-to-answer paths; the cases below provide provenance and metric context.

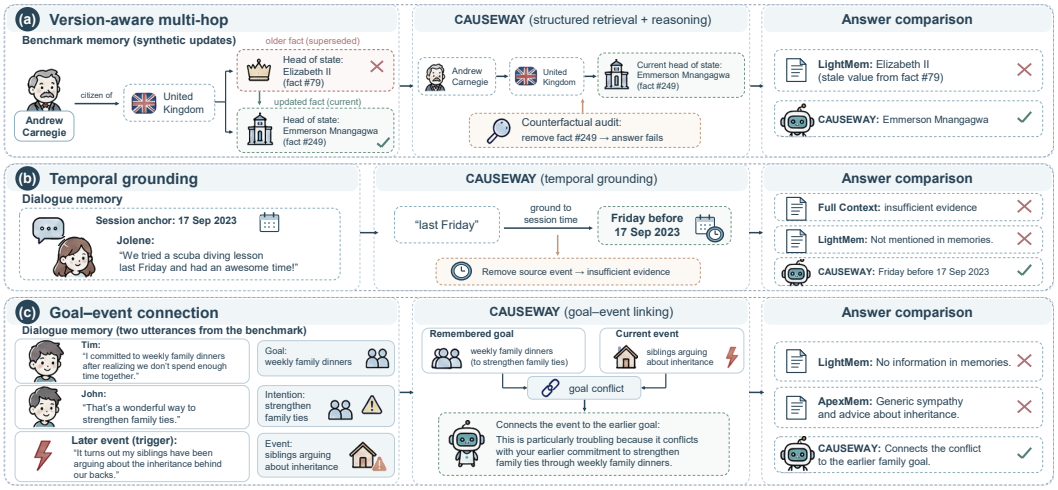

Figure 6: Selected source-level cases: (a) a superseded assignment changes a multi-hop dependency; (b) a session date anchors a relative-time expression; (c) an earlier goal constrains a later response. These are illustrative, non-random examples.

Panel (a) uses `factconsolidation_mh_6k_no14` in the `Conflict_Resolution:factconsolidation_mh_6k:row0` context: facts 79, 249, and 367 link a superseded leader assignment to the current answer. These assignments are synthetic benchmark facts. Panel (b) is `locomo:conv-48:q76`, whose September 17, 2023 session anchors "last Friday." Panel (c) is `locomo_plus:official_cognitive:0032`, connecting family dinners to the earlier goal of maintaining family ties. The panels are selected, non-random illustrations.

The additional LoCoMo cases are the lexicographically first available success/failure within two types with a saved source excerpt.

Anchoring relative time. For `locomo:conv-26:q1`, a source dated May 8, 2023 says that the sunrise was painted "last year." Both CAUSEWAY and LightMem answer 2022 and receive score 1. Both systems recover the same time anchor.

Zero F1 is not a semantic failure label. For `locomo:conv-26:q68`, the question asks how long Melanie has practiced art; the reference is "Since 2016," while both systems answer "Seven

years." A September 2023 source explicitly says seven years. The recorded zero score reflects a mismatch between duration and start-year surface forms.

Event counting after fallback. For `locomo:conv-26:q40`, the question asks how many beach visits occurred in 2023. Saved sources include a July family beach-camping mention and a later "recently" visit. The graph-stage audit abstains, while v3 emits "once or twice a year," receiving zero against the reference 2; LightMem answers "Twice." This illustrates a failure of event counting and the distinction between an audited graph answer and a post-audit fallback output.

Related sentiment is not the relevant historical constraint. On Plus item `0002`, the matched judge evidence connects a child's trampoline injury to replacement of the trampoline with a garden. The full route recalls that relation, whereas the semantic-flat route discusses relief after finishing a screenplay. The latter response resembles the trigger's emotional language but does not connect to the designated cue. This route-level illustration uses the saved judge evidence and prediction. The canonical question field comes from a converted export rather than a complete runtime prompt snapshot, so reader-visible cue and speaker fields cannot be reconstructed from it.

A.11 Reproducibility and Resource Accounting

The original audit reconstructs 13,210 canonical main records: 13,192 scored and 18 explicit missing slots. The MAGMA supplement adds 1,321 scored records (600 LoCoMo, 401 Plus, and 320 MAB), bringing combined coverage to 14,531 records, 14,513 scored, with the same 18 missing slots. Offline reanalysis reproduces the main aggregates and reruns the bundled LoCoMo scorer on the original 6,000 main-method outputs to floating-point tolerance, without model calls; the separately scored 600 MAGMA outputs bring deterministic LoCoMo coverage to 6,600. The audit provides input hashes, unrounded aggregates, and analysis code; newly computed intervals are marked exploratory.

The controlled mechanism studies of Appendix A.8 preserve protocol, snapshot, code and request hashes, raw responses, and atomic per-item checkpoints. Older exports lack some per-record model identifiers, prompt hashes, code commits, or cache provenance, limiting request-level reproducibility for those runs. The following profiles separate construction and query stages with explicit denominators. Tokens use provider-reported input/output usage; other exporters' `cache_tokens` fields are excluded from billed-cache interpretations where their semantics are undocumented.

A.11.1 Query-Resource Accounting

Evaluation and resource-measurement scope. Table 24 extends the main-text summary (Table 3) to all context and retrieval baselines, with separate input and output usage, combining accuracy from the main evaluation with query-stage LLM resource measurements on the same frozen 401-item Cognitive set.

For Causeway, the retained checkpoints record 3,333 API attempts, including 19 errors, with 3,141,715 input and 453,958 output tokens. Dense RAG records 403 requests, including two errors, with 1,862,577 input and 43,563 output tokens. ApexMem records 2,823 network requests, including five errors, with 2,270,112 input and 173,064 output tokens. Its 14 response-cache hits are excluded from network usage.

MAGMA's 401-question reader profile records 418 API attempts, including 17 connection failures, with 1,022,634 input and 4,747 output tokens (1,027,381 total). It uses the saved retrieval contexts, native prompt and answer postprocessing, MiMo v2.5 at temperature zero with thinking disabled, and a 3,000-token output ceiling. The recorded 120,640 cached-input tokens are already included in input usage; missing usage from failed requests is unknown. Its mean is 2.562k tokens per query, 6.57% of Full Context and 28.57% of Causeway. Per-attempt usage, request hashes, and response identifiers support the totals.

Denominators, caches, and retries. The denominator is the fixed 401 questions, including refusals. Tokens sum reported usage from successful network responses; token usage for failed requests is not available. API-attempt counts include the recorded errors. Successful network-call counts are 401 for Dense RAG, 401 for LightMem, 401 for MAGMA, 2,818 for ApexMem, and 3,314 for Causeway. For ApexMem, per-question logs are reconciled with the saved call, token, and cache-hit totals; Causeway accounting uses per-question checkpoint totals. Application-cache hits are

Table 24: Main-evaluation accuracy and query-stage LLM usage on LoCoMo-Plus. Tokens are thousands per question; attempts include failures. Construction and judging are excluded.

Method	Accuracy (%)	Input (k tok)	Output (k tok)	Total (k tok)	API attempts per query
Context and retrieval baselines					
Full Context	34.41	38.861	0.113	38.974	1.00
Truncated Context	33.92	38.861	0.112	38.974	1.00
BM25	7.23	5.003	0.109	5.112	1.00
Dense RAG	8.48	4.645	0.109	4.753	1.00
Hybrid RAG	6.98	5.010	0.109	5.119	1.01
Memory systems					
LightMem	57.61	0.773	0.397	1.170	1.56
MAGMA	17.71	2.550	0.012	2.562	1.04
ApexMem*	82.29	5.661	0.432	6.093	7.04
CAUSEWAY	**87.78**	7.835	1.132	8.967	8.31

excluded; provider-reported cached-input tokens remain part of input usage. Failed-request usage is unavailable, and tokens are not converted to dollar cost.

Additional context and retrieval baselines. For BM25, Hybrid RAG, Truncated Context, and Full Context, every retained checkpoint prediction matches the corresponding main-evaluation prediction on all 401 item IDs. Each method has 401 successful answer calls with reported input and output usage. BM25 records 403 attempts and 2,006,018 input plus 43,887 output tokens; Hybrid RAG records 406 attempts and 2,009,033 input plus 43,868 output tokens. Truncated Context and Full Context each record 403 attempts and 15,583,404 input tokens, with 45,062 and 45,143 output tokens, respectively. On all 401 items, both context baselines retain the same ordered evidence and report no truncation under the configured local evidence-length estimator. This explains the identical input totals; the provider's token counts, rather than that local estimator, are used for resource accounting.

Partial A-Mem and Mem0 telemetry. The main-evaluation accuracies are 14.71% for A-Mem and 8.73% for Mem0 on all 401 questions. To associate resource records with these results, we match the question, ordered evidence, model, and final answer between retained checkpoints and raw successful-call logs. This yields 74 A-Mem and 160 Mem0 answer calls with provider-reported usage (Table 25). These partial means cover matched final-answer calls only, excluding unlinked calls, retries, and construction. They are omitted from the full-set comparison in Table 24.

Table 25: Partial final-answer telemetry for A-Mem and Mem0. Accuracy uses the full 401-item main evaluation; token means use only the explicitly matched calls in the coverage column. Tokens are in thousands per matched call. Coverage is partial and limited to the final-answer stage.

Method	**Accuracy (%)**	**Coverage**	**Input**	**Output**	**Total**
Matched final-answer calls only					
A-Mem	14.71	74/401	3.678	0.139	3.817
Mem0	8.73	160/401	0.593	0.359	0.952

Latency. Wall-clock query latency is elapsed time from query entry to final answer, including retrieval, local processing, API waiting, and retries, but excluding prior memory construction and offline benchmark judging. Dense RAG has query-boundary timestamps. CAUSEWAY and ApexMem provide API durations without equivalent query-boundary timestamps. LightMem's and MAGMA's timers start from fixed retrieval evidence and therefore cover the answer stage. Table 26 reports successful API durations for all five methods; end-to-end wall-clock statistics are available only for Dense RAG, at 4.48 s mean, 3.98 s p50, and 6.34 s p95. Service conditions differ across profiles; timings are observational.

Table 26: Recorded timing on the same 401 Plus queries. API time is the mean sum of successful network-request durations per query, in seconds, and is not end-to-end latency. Only Dense RAG has query-boundary timestamps, with 4.48 s mean, 3.98 s p50, and 6.34 s p95 wall-clock time; the other profiles lack equivalent timers.

Method	API time (s)
Dense RAG	3.99
LightMem	17.52
MAGMA	12.58
ApexMem*	17.81
CAUSEWAY	44.30

Construction and amortization. The query profile excludes prebuilt memory and index construction. For a fixed history workload with Q queries, amortized token usage is

$$T_{\text{amortized}} = \frac{T_{\text{build}} + \sum_{q=1}^{Q} T_{\text{query},q}}{Q}.$$

Construction should be counted once per history and memory snapshot; question-specific memory updates belong to the corresponding query workload. Incomplete Plus construction logs, including missing MAGMA build telemetry, prevent an amortized total or break-even estimate. Embedding/indexing resources and complete end-to-end timers are also unavailable.

A.11.2 LIGHTMEM CONSTRUCTION AND ANSWER COSTS

LoCoMo. The LoCoMo profile uses all 600 frozen LightMem answer records and the construction records for their ten corresponding conversation histories. Source-artifact hashes identify the answer records, and each history's construction is counted once. These records contain 1,242,055 input tokens, 2,154,940 output tokens, and 806 recorded construction LLM calls. The 600 successful answer completions contain another 1,615,423 input and 579,258 output tokens. Table 27 separates answer-only usage from construction amortized over this fixed 600-query workload.

Table 27: LightMem construction and answer usage. Tokens are means per question, in thousands. LoCoMo amortization counts the ten corresponding history builds once and divides their usage by 600. MAB rows cover only the 200 generic-route questions and exclude the preceding evidence-selector LLM calls.

Benchmark	Accounting scope	n	Input	Output	Total
LoCoMo	Answer stage	600	2.692	0.965	3.658
	Build amortization only	600	2.070	3.592	5.662
	Build + answer	600	4.762	4.557	9.319
MAB	Native build stage	200	3.772	2.389	6.161
	Native answer stage	200	1.073	0.008	1.081
	Recorded native stages	200	4.845	2.397	7.242

MemoryAgentBench. All 320 checkpoint predictions match the frozen LightMem result. Of these, 200 use the generic route, which retains native construction statistics and answer usage; the remaining 120 use conflict, infbench, or recsys routes. The 200 native construction scopes are question-specific and distinct, so their build cost belongs to each query rather than to a shared-history amortization. They record 754,440 input tokens, 477,755 output tokens, and 574 construction LLM calls. Their successful answer completions add 214,606 input and 1,594 output tokens. The native-ingestion call counter agrees with the construction-summary counter for every generic item; the two counters are not added. The preceding evidence selector is outside these usage fields, and the other routes lack equivalent token records. The reported MAB profile is restricted to these 200 questions.

Calls, timing, and benchmark boundaries. The call counters omit unrecorded transport retries. LoCoMo query time averages 27.88 seconds over 600 questions, excluding prior construction. MAB

execution time averages 44.53 seconds over 320 questions, including route-specific work and in-query construction. These timings differ in scope and service conditions. LLM-token profiles exclude embedding computation and offline judging.