EVSTATEMEM: GOVERNING CONVERSATIONAL MEMORY THROUGH STATE AND EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Information in long-term conversations can be revised, withdrawn, or restricted in scope. Even when a memory is relevant to the current question, an agent can violate user intent by using outdated information, exceeding its permitted scope, or bringing up content that should not be invoked. Conversational memory systems therefore need to track information validity over time and constrain its use according to user instructions and the current context. We call this requirement governed use of memory and introduce EVStateMem to address it. The system separately manages source records, versioned states, and answer evidence, retaining statement provenance and the evidence supporting state changes. A derived Current View organizes current states, conflicts, and suppressions; answer-time evidence selection checks source availability, scope, and lifecycle status. We evaluate conversational memory question answering on LoCoMo and LongMemEval, and use TimlMemBench's nine behavioral diagnostics to assess adherence to memory-update and use constraints. EVStateMem achieves competitive question-answering results on both benchmarks and the highest scores among the compared configurations on seven of TimlMemBench's nine behavioral axes, including revocation, replacement, scope control, and selective activation.

1 INTRODUCTION

A conversational agent can retrieve a relevant statement that has been corrected, withdrawn, or restricted to another context. Remembering the statement is useful for a historical question, while using it as the current value can produce an incorrect answer. A persistent memory system therefore needs to determine what an update changes and which stored information a later response may expose. We use *memory governance* to mean the rules that determine which information may be stored, revised, and used in a response.

Existing systems already support important parts of this process. Mem0 extracts facts and chooses update operations; A-MEM organizes and revises linked notes; Zep tracks changing relationships and their temporal validity (Chhikara et al., 2025; Xu et al., 2025b; Rasmussen et al., 2025). The remaining design question concerns the connection between these operations. Replacing a value must preserve the evidence needed for historical reads, and a retrieved record must still satisfy its source and scope constraints at answer time. Relevance ranking alone cannot establish those conditions.

EVStateMem makes these conditions explicit across the write and read paths. A *source record* stores an utterance with its speaker, scope, and time information. A *derived memory* stores searchable content linked to those records. A *relation state* records a subject–relation–value assertion, its validity interval, and its supporting sources. These objects separate what was said from the state derived from it. Source references connect both to the bounded *evidence set* selected for a particular answer.

Versioned state and question-time selection implement memory governance together. Explicit assertions, replacements, and withdrawals determine which versions remain active. *Current View* (CV) is a derived read model that groups current records, their evidence, unresolved conflicts, and suppressed records. Eligible current-value requests can use this view; historical questions read state versions, and general questions retrieve complementary evidence. Each path checks scope, lifecycle status, and supporting sources before generation. Figure 1 shows how these paths connect state maintenance to evidence use.

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.

We contribute a memory architecture that separates recorded statements, versioned state, and answer evidence while retaining their source links. Its update-and-read procedure preserves history and rechecks evidence eligibility across reading paths. We evaluate both conversational question answering and instructed memory behavior.

LoCoMo measures recall and reasoning over conversations, while LongMemEval also tests knowledge updates, temporal reasoning, and abstention (Maharana et al., 2024; Wu et al., 2025). TimlMem-Bench (Zhou et al., 2026a) adds targeted tests of withdrawal, scope isolation, source recovery, disclosure boundaries, and permitted memory use. This separates recovering an answer from following instructions about how remembered information should be handled. With GPT-5.6-terra, EVStateMem leads sixteen local LoCoMo baselines on single-hop and multi-hop F1 and reaches 81.60% on LongMemEval S-cleaned. The behavior evaluation reports strengths and remaining errors by operation.

2 Problem formulation

Let $\mathcal{D} = (d_1, \ldots, d_n)$ be an ordered conversation history containing n source records. Each record has the form

$$d_i = (\mathrm{id}_i, a_i, \sigma_i, \tau_i, x_i), \tag{1}$$

where id_i is a persistent identifier, a_i is the speaker and message role, σ_i is the access scope, τ_i is the source time, and x_i is the original text. A scope identifies the conversation, user, or other authorized context to which a record belongs. The source time may be an instant, a calendar date, or unknown. The time at which the system imports a record is stored separately.

A question consists of text q, an authorized scope set Σ_q, and a reference time τ_q when available. The system selects an ordered evidence set $\mathcal{E}_q$ of at most B items and supplies it to an answer model A:

$$\hat{y} = A(q, \mathcal{E}_q), \qquad |\mathcal{E}_q| \le B. \tag{2}$$

Here $\hat{y}$ is the generated answer, and B bounds the number of evidence items. The model's context limit imposes a separate token constraint. The memory procedure receives the conversation and question; reference answers belong to the evaluator.

Two distinctions determine which evidence is appropriate. First, a stored statement records what a speaker said, including statements that are later corrected. A state view selects which supported value applies to a particular time and scope. Second, the content used to find a memory may be shorter than the content needed to answer from it. EVStateMem therefore retains source references through extraction, state updates, retrieval, and final evidence selection.

We call a selected item *eligible* when its scope is authorized, its lifecycle status permits the requested read, and its required source links resolve. Lifecycle status records whether a memory is active, archived, superseded, or deleted. Current queries use the currently valid version, while historical queries select the version valid at the specified time. When historical access is permitted, superseded evidence remains usable, but sources that have been deleted or fall outside the scope authorized for the query are always excluded.

3 The EVStateMem method

3.1 Preserve sources and derive searchable memory

The write procedure first stores a source record, then derives memory items linked to it. Short-term items retain recent conversational context. Episodic items describe particular events, and semantic items describe reusable facts or preferences. Extraction produces content, retrieval summaries, and source references. Consolidation forms episodic memories from short-term items; semantic refinement derives reusable content from episodic items. Maintenance deduplicates, revises, and archives memory items. Summaries help rank candidates; linked source text supplies the wording and attribution needed during evidence selection.

Separating source records from derived memory gives updates a stable evidence basis. Re-indexing changes a memory's searchable form while retaining its underlying utterance. Supersession replaces

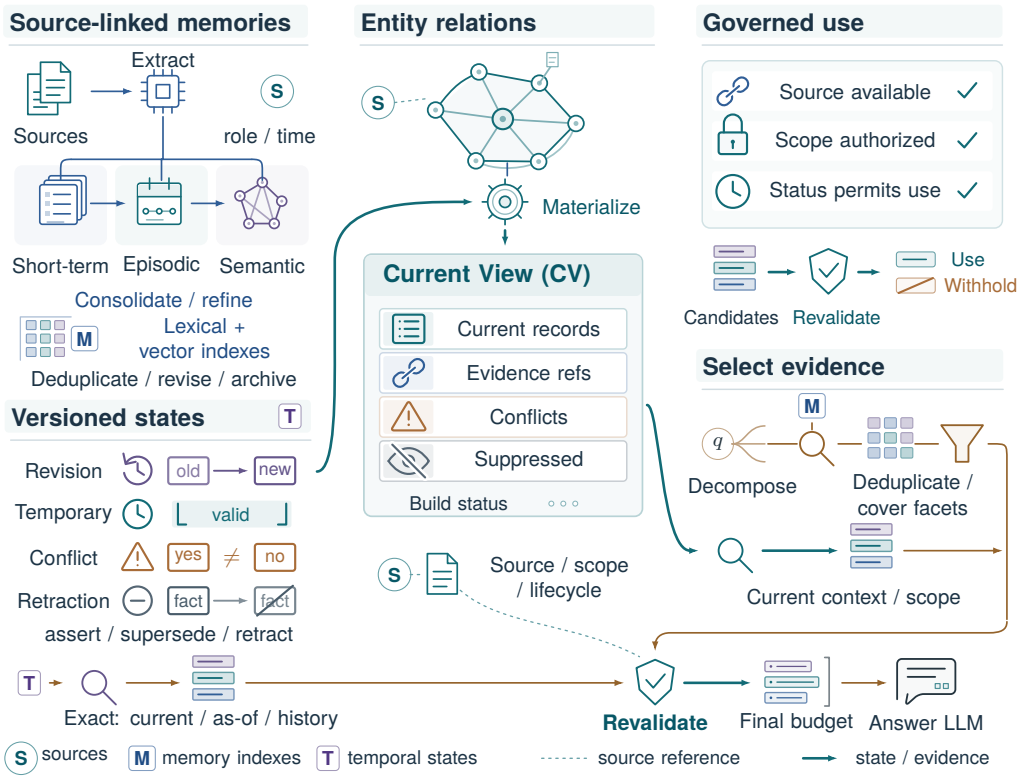

Figure 1: EVStateMem overview. Source-linked memories support retrieval; versioned states and entity relations supply the derived Current View (CV), which separates current records, evidence references, conflicts, and suppressions. General retrieval, current-view context, and exact state reads share answer-time source, scope, and lifecycle checks. The upper-right inset details this revalidation step: only eligible evidence can proceed to the final budget and answer generation. Repeated S, M, and T markers identify sources, memory indexes, and temporal states; q denotes the question. Dotted teal lines denote source references. Record contents and status marks are schematic.

an active value while preserving sources for earlier versions. Generation traces are stored separately, so inspection of an answer does not turn that answer into a new user assertion.

Speaker and time fields constrain extraction. The implementation distinguishes user, assistant, system, and tool messages. An assistant's description of a user does not independently authorize a durable user-state update. When a source supplies only a date, state processing keeps date precision. Import time supplies no missing event date. These checks operate on persisted speaker and time fields.

3.2 Update relation states with explicit operations

A relation state is a source-supported assertion about a subject. We represent a state record as

$$s = (u, r, v, \sigma, I, L, \omega), \qquad I = [t_{\text{start}}, t_{\text{end}}), \tag{3}$$

where u is the subject, r the relation identifier, v its value, σ its scope, I its validity interval, L its supporting source references, and ω its state status. An open end is written as $t_{\text{end}} = \infty$. Interval comparisons require compatible time precision. A relation definition specifies the subject type, value type, cardinality, and aliases. Cardinality states whether the relation admits one value or a set of values for the same subject and scope.

The state writer supports three operations. `assert` adds a supported value. `supersede` closes explicitly identified earlier versions and appends a replacement. `retract` records a withdrawal and preserves its source evidence. All operation targets must match the subject, relation, and scope of the update. For single-valued assertions with overlapping validity intervals, mutually exclusive values,

and no clear evidence of revision, the system preserves both sources and marks the conflict for user review. Set-valued assertions retain separate members and their evidence.

Temporary state has a bounded interval. It can overlay a durable value during that interval; after expiry, the durable value becomes visible again. Exact replays of a previously applied update are idempotent: they leave the resulting state unchanged. Ambiguous timestamps, incompatible date precision, and unsupported destructive changes are rejected or marked for review. The source utterance remains available even when state derivation fails.

3.3 Materialize current state and retain historical access

Current View materializes temporal states and entity-relation records into a read model. Current records, evidence references, unresolved conflicts, and suppressions remain distinct, so a stored record need not become active context. Build status records the latest rebuild and its outcome. When enabled, the context builder selects up to three eligible records for supported current-value requests within the authorized scope. Historical requests read the relevant versions or source records directly.

Reading begins with scope authorization and query analysis. The system identifies an explicit subject, relation, and time constraint when the question supports that binding. Exact state queries select current values, values at a specified time, or state history, together with supporting sources. Current-value requests can also use Current View.

3.4 Collect complementary evidence for the question

For general questions, lexical and embedding retrieval produce an initial candidate pool. Ranking accounts for query relevance alongside memory status and scope. A configured reranker can reorder candidates within this pool. Retrieval summaries are used as search hints. Evidence projection resolves the linked source text and retains speaker attribution before the final answer request.

Questions involving shared properties, comparisons, lists, or entity links can require several searches. The planner creates at most three additional queries in one stage. Rule-based templates cover explicit question forms and carry their explicit subjects and constraints into the branches. A supported entity-link query can use a uniquely identified source entity in the next stage, with a maximum of three stages. An optional model-based planner proposes additional retrieval wording when enabled. Its prompt requests preservation of subject, ownership, and time constraints. The parser rejects newly introduced named, numerical, temporal, and quoted anchors absent from the question and supplied evidence hints.

Let $q_0 = q$ be the original question and $q_1, \ldots, q_m$ the accepted branches in one stage, with $m \leq 3$. Candidate collection and evidence selection take the form

$$\mathcal{C}_q = \bigcup_{j=0}^{m} \mathrm{Retrieve}(q_j, \Sigma_q), \qquad \mathcal{E}_q = \mathrm{Select}(q, \mathcal{C}_q, B). \tag{4}$$

Here $\mathrm{Retrieve}$ returns candidates for a query and its authorized scopes, $\mathcal{C}_q$ is the combined pool, and Select returns ordered evidence under budget B. The union removes duplicate memory identities. Every branch uses the same authorized scopes. Branching changes the search wording and candidate coverage; the final selection retains the original evidence budget.

The coverage planner prefers evidence that adds a requested facet, such as a second subject, another event, or a time anchor. It uses relevance and textual overlap to order the remaining candidates. The current message-linked implementation keeps the highest-ranked anchor and then prefers new query-term coverage and distinct source messages. This intermediate ordering examines at most 200 candidates and selects at most 100 items. Final factual-answer projection imposes its own 20-item cap. Coverage signals guide selection; the answer model still has to determine whether the selected text supports the requested conclusion.

3.5 Revalidate sources and generate the answer

Before generation, EVStateMem checks each selected item's lifecycle status, authorized scope, and source availability again. State-derived evidence must resolve to an eligible source; a graph entry

alone cannot authorize prompt inclusion. A missing or withdrawn supporting record blocks the corresponding projection. These checks also apply when evidence was found through a query branch or a relation-state lookup.

The answer request contains the question and the selected source-linked evidence. Factual answer generation is separated from conversational style context. Structured factual routes retain the evidence references associated with emitted values. The trace records retrieval branches, selected items, state-read decisions, and the provider call. Inspecting a trace reveals which sources entered the request and which operation selected them. Changing an answer provider preserves these memory interfaces.

This separation links state management to answer construction. An update determines a record's validity; Current View exposes eligible current records; source revalidation determines whether selected content can enter the answer request. General retrieval and exact state reads retain their own access paths and use the same final evidence checks.

4 EVALUATION

4.1 EVALUATION SETUP

We evaluate two aspects of conversational memory. LoCoMo and LongMemEval test recovery and interpretation of conversational evidence. TimlMemBench tests adherence to memory-update and use instructions, including when relevant information should be withheld.

F1 measures predicted–reference token overlap; B1 denotes BLEU-1, unigram precision with a length penalty. Semantic Judge accuracy J measures acceptance of saved answers. LongMemEval reports judged answer accuracy; TimlMemBench reports behavior acceptance by axis. Appendix A provides full metrics and protocols.

4.2 CONVERSATIONAL MEMORY QA

LoCoMo. We use the ten-conversation LoCoMo release (Maharana et al., 2024), covering 1,986 questions. Memory is built before questions are submitted; gold answers are used only for scoring. Table 1 reports the four non-adversarial categories, with 1,540 semantic judgments per run.

Table 1: LoCoMo results (%): F1/B1 with GPT-5.6-terra. Bold and underline mark the best and second-best distinct scores per column; ties share the same mark. Full metrics appear in Appendix A.3.

		Single-hop		Multi-hop		Open		Temporal	
Method	Overall F1	F1	B1	F1	B1	F1	B1	F1	B1
A-MEM	45.00	49.77	43.47	33.66	26.95	26.47	22.94	48.01	42.14
Graphiti	41.29	45.61	39.60	29.52	24.28	29.63	24.08	43.82	38.42
LangMem†	14.51	15.83	12.85	14.20	11.54	12.50	11.07	11.93	10.43
LightMem	<u>56.36</u>	58.17	50.32	42.40	34.96	<u>33.74</u>	27.95	**70.67**	**64.66**
Mem0	24.96	30.43	21.67	26.19	18.98	18.45	14.34	11.50	8.32
MemoryBank-summary†	45.41	45.54	38.01	40.86	32.58	27.67	23.30	54.37	47.25
MemoryOS	48.85	52.86	46.18	38.97	32.12	32.81	<u>28.20</u>	51.80	45.17
MemPalace	49.86	54.57	47.76	34.17	27.98	32.70	27.31	56.42	52.12
memU†	33.11	34.70	29.28	30.92	24.89	22.18	18.74	34.15	29.52
Omni-SimpleMem	51.69	54.47	46.70	<u>44.35</u>	<u>37.22</u>	**34.05**	**29.07**	56.13	49.33
OpenClaw	45.93	53.67	46.27	38.31	31.03	27.27	21.90	37.93	32.51
PowerMem	48.99	51.64	44.75	32.73	26.63	30.39	24.93	61.91	54.91
PropMem	56.17	<u>61.18</u>	<u>53.99</u>	39.34	32.28	32.77	27.07	<u>64.85</u>	<u>59.11</u>
RAG+BGE-M3	52.79	58.52	51.06	34.98	28.51	27.67	22.55	60.95	55.83
ReMe†	21.71	21.34	12.68	24.16	15.81	9.33	5.51	24.24	15.18
SimpleMem†	53.03	55.22	47.46	43.06	34.94	27.61	24.39	63.67	57.62
EVStateMem	**61.34**	**69.11**	**62.13**	**52.30**	**45.57**	30.03	25.48	58.32	53.01

† identifies the Terra configurations described in Appendix A.3.

Both EVStateMem runs use a primary retrieval limit of 20 items. Appendix A.3 records the full baseline settings.

Table 1 reports local Terra runs for sixteen baselines and EVStateMem. MemoryBank uses its summary component. Appendix A.3 provides semantic Judge scores and GPT-4o results. Retrieval budgets and answer prompts vary.

With Terra, EVStateMem reaches 69.11 single-hop and 52.30 multi-hop F1. Across the full local comparison in Table 6, these exceed the strongest baselines by 7.93 and 7.95 percentage points; B1 also leads both categories.

Question-weighted J is 70.78% with Terra. Appendix A.3 retains semantic-score comparisons and separately reports adversarial results with their distinct protocols.

LongMemEval. LongMemEval tests recall, preferences, temporal reasoning, and knowledge updates across conversation histories (Wu et al., 2025). Table 2 reports the Qwen group on all 500 S-cleaned questions, including eight MemForest-reported configurations (Chen et al., 2026). Further LongMemEval S-cleaned experiments and the complete table appear in Appendix A.1.

Table 2: LongMemEval S-cleaned accuracy (%). SS-U/A/P: single-session user/assistant/preference; Multi: multi-session; Temp: temporal; Update: knowledge update. Bold and underline mark the best and second-best distinct scores within each model group.

Method	Overall	SS-U	SS-A	SS-P	Multi	Temp	Update
Answer: **Qwen3-4B-Instruct-2507**						Judge: DeepSeek-V4-Flash	
MemForest-Planner	**72.60**	**95.71**	83.93	<u>70.00</u>	59.40	**70.68**	70.51
MemForest-Embed	69.40	88.57	83.93	53.33	60.90	**70.68**	60.26
EverMemOS	61.20	84.29	69.64	66.67	55.64	48.12	64.10
LightMem	68.80	<u>94.29</u>	35.71	**76.67**	**65.41**	64.66	**79.49**
MemoryOS	64.20	87.14	<u>89.29</u>	46.67	57.14	60.90	50.00
MemPalace	60.00	91.43	33.93	36.67	<u>63.91</u>	52.63	65.38
Mem0	44.80	82.86	26.79	33.33	42.11	38.35	43.59
Zep Local	69.60	<u>94.29</u>	**98.21**	53.33	59.40	54.89	<u>75.64</u>
EVStateMem	<u>71.40</u>	92.86	**98.21**	53.33	51.13	<u>68.42</u>	**79.49**

With Qwen, it reaches 71.40% (357/500), second to MemForest-Planner's 72.60%. It matches the highest assistant-side (98.21%) and knowledge-update scores (79.49%); preference, multi-session, and temporal scores remain below the respective leaders. These comparisons evaluate the recorded reader and memory configurations.

4.3 Memory-state and use diagnostics

A correct answer to an updated fact tests one part of memory governance. An agent must also follow instructions to withdraw information, keep contexts separate, recover sources, and avoid unrequested use. TimlMemBench (Zhou et al., 2026a) provides targeted behavior cases for these decisions. Its 208 human-reviewed synthetic dialogue samples contain 1,566 sessions. We evaluate EVStateMem on all 962 Chinese and 859 English behavior cases, ingesting each sample session by session; gold states and rubrics are available only to the judge.

The nine axes distinguish state maintenance from memory use. Write tests storage decisions; Revoke and Supersede test behavior after withdrawal and replacement; Preserve tests retention across distractions. Scope tests context isolation, Provenance tests source recovery, and Boundary tests disclosure constraints. Activation tests relevant use and suppression of irrelevant or forbidden memories; Proactive tests restraint against unsolicited use. Scores are semantic-judge acceptance rates on final responses, with overlapping case subsets.

Table 3 pools Chinese and English accepted-case counts separately for each axis and includes all twelve published baselines. Language-specific results and aggregation details appear in Appendix A.5; question-answering controls and run settings appear in Appendix A.4.

Table 3: TimlMemBench behavior scores (%), pooling Chinese and English cases within each axis. Baselines come from Zhou et al. (2026a), with GPT-5.4-mini answering and judging; EVStateMem uses that gateway-reported model label. Bold and underline mark the best and second-best distinct scores per column; ties share the same mark. Proactive measures restraint.

System	Write	Revoke	Supersede	Preserve	Scope	Provenance	Boundary	Activation	Proactive
Mem0	43.9	47.8	51.4	45.6	48.0	44.7	52.9	54.1	63.3
MemoryOS	**67.2**	<u>68.2</u>	67.0	61.9	62.6	**72.4**	64.9	66.7	66.8
MemoryBank	51.9	66.1	<u>67.6</u>	63.0	66.0	59.2	<u>70.4</u>	<u>71.9</u>	<u>72.1</u>
MemPalace	31.2	39.8	36.8	38.9	45.2	44.7	34.3	42.5	50.5
Graphiti	48.7	48.7	37.5	53.1	57.0	53.9	38.2	53.7	59.4
Zep	29.1	38.4	34.6	48.8	49.7	33.8	44.9	47.1	58.7
LightMem (SLM)	39.7	44.9	37.5	52.5	54.8	47.4	45.5	48.4	58.0
ReMe	36.5	40.7	27.3	48.8	53.9	34.2	48.7	44.4	59.7
PowerMem	36.0	54.7	44.8	53.4	55.6	45.2	54.5	59.2	68.9
A-MEM	<u>59.8</u>	66.6	67.3	<u>63.3</u>	<u>66.3</u>	<u>68.0</u>	66.3	67.6	70.3
SimpleMem	35.4	43.2	28.3	49.6	54.5	42.5	37.5	44.9	64.0
TriMem	41.3	54.0	41.9	54.7	54.5	50.4	53.0	54.7	65.0
EVStateMem	52.4	**76.2**	**70.5**	**74.8**	**76.7**	62.7	**77.3**	**78.0**	**86.6**

EVStateMem has the largest reported score on seven of the nine axes, including 86.6% Proactive and 78.0% Activation accuracy. Revoke and Supersede reach 76.2% and 70.5%. Write is 52.4% and Provenance is 62.7%, below MemoryOS at 67.2% and 72.4%. The resulting profile locates stronger restraint and state-update behavior alongside remaining errors in retaining and recovering evidence. These scores evaluate complete memory-and-answering configurations.

4.4 Component analysis

The TimlMemBench ablation configurations differ only in whether the target components are enabled. All other code, prompts, LLM configurations, and evaluation procedures remain unchanged, and all configurations are evaluated on the same test cases.

Removing versioned state disables its maintenance and reading while retaining source memories, standard retrieval, and Current View built from entity-relation records. Removing Current View keeps versioned state storage but stops using that view during answering.

In TimlMemBench removal experiments, removing versioned state reduces Supersede scores by 8.6 percentage points in both languages; removing Current View reduces them by 4.5 points in Chinese and 4.3 in English. Removing final eligibility rechecking alone yields score decreases of 0–1.6 points across axes; its role in blocking evidence invalidated after retrieval is further demonstrated by the controlled test in Appendix A.11. Removing all three reduces every axis in both languages. In the cumulative LoCoMo comparison, the full configuration improves multi-hop F1 by 7.06 points over configuration D, with temporal F1 decreasing by 3.00 points. Appendix A.8 gives configurations and full scores; further component analysis appears in Appendix A.7.

A controlled post-selection revocation test isolates final source rechecking. After selection, revocation changes the persisted record to `deleted` while its cached status remains `active`. With rechecking enabled, the revoked evidence is excluded from the Reader request; without it, the evidence remains. Unchanged evidence is retained in both conditions. This demonstrates enforcement before model invocation when source eligibility changes after retrieval (Appendix A.11). Additional eligibility and efficiency measurements appear in Appendix A.9.

5 Related work

Persistent conversational memory. Persistent conversational memory addresses the mismatch between an accumulating interaction history and a bounded model context. MemoryBank retrieves and updates long-term memories with a forgetting mechanism (Zhong et al., 2024). MemGPT manages transfers between memory tiers, treating context construction as an explicit part of the

Table 4: TimlMemBench component removals (%). Bold and underline mark the highest and second-highest distinct scores per column within each language; ties share the same mark.

Configuration	Write	Revoke	Super.	Preserve	Scope	Prov.	Bound.	Activ.	Proact.
ZH									
Full system	**50.9**	**73.9**	**68.5**	**69.4**	**69.0**	**60.9**	**78.5**	**75.7**	**82.7**
w/o versioned state	47.3	68.1	59.9	65.7	67.0	57.8	74.9	71.1	80.6
w/o Current View	<u>49.1</u>	70.7	64.0	<u>66.7</u>	<u>68.0</u>	<u>59.4</u>	77.0	73.0	<u>81.6</u>
w/o final recheck	**50.9**	<u>72.9</u>	<u>68.0</u>	**69.4**	<u>68.0</u>	<u>59.4</u>	<u>78.1</u>	<u>75.0</u>	<u>81.6</u>
w/o all three	43.6	63.8	55.9	62.0	64.0	54.7	71.0	67.1	77.6
EN									
Full system	**53.0**	**77.9**	**75.3**	**77.0**	**79.7**	**63.4**	**73.5**	**80.5**	**88.6**
w/o versioned state	49.3	71.9	66.7	72.1	77.0	59.1	70.1	76.0	87.0
w/o Current View	<u>50.7</u>	75.1	<u>71.0</u>	<u>74.0</u>	<u>78.1</u>	61.0	72.1	78.1	<u>88.1</u>
w/o final recheck	**53.0**	<u>77.1</u>	**75.3**	**77.0**	<u>78.1</u>	<u>62.2</u>	<u>72.8</u>	<u>80.0</u>	<u>88.1</u>
w/o all three	46.3	67.9	62.4	67.9	73.0	56.1	66.7	71.9	83.8

agent's execution (Packer et al., 2023). MemoryOS organizes short-term, mid-term, and long-term personal memory through storage and update procedures (Kang et al., 2025). For conversational agents, the retained information also carries conditions: a preference may be temporary, a statement may be corrected, and a memory may belong to a different interaction scope. EVStateMem carries these conditions through the write and read paths, connecting persistent storage to the eligibility of evidence used in an answer.

Structured updates and memory organization. Memory systems already support substantial reorganization and revision. Mem0 compares extracted facts with existing memories and selects addition, update, deletion, or no change (Chhikara et al., 2025). A-MEM creates structured notes, links them to related memories, and revises existing attributes and contextual descriptions as new memories arrive (Xu et al., 2025b). EVStateMem makes the targets and consequences of state updates explicit. An operation identifies a subject, relation, and scope; supersession retains the earlier version and its sources, while unresolved conflicts remain visible to state processing.

Temporal state, provenance, and governed use. Zep's Graphiti is a close reference for source-linked temporal memory. It separates episodes from extracted entities and relationships, records both ingestion time and fact validity, and invalidates conflicting edges while retaining historical relationships (Rasmussen et al., 2025). MemForest organizes canonical facts in scoped temporal trees and refreshes affected summaries locally (Chen et al., 2026). EVStateMem develops the connection between such state records and response-time use. Its Current View groups current records, evidence references, conflicts, and suppressions into a derived read model; historical requests retain access to the relevant versions. Source availability, authorized scope, and lifecycle status are checked during evidence selection. These checks distinguish historical support, current validity, and eligibility for use, including cases where retained information should stay outside the answer context.

Recent work makes evidence and lifecycle handling explicit. Eywa preserves source evidence before deriving facts (Joshi, 2026); MemoryLACE links memories through merge, supersession, and contradiction relations (Yacoubi et al., 2026). A-TMA constructs state-labeled evidence packets (Shi et al., 2026), TEPA revokes stale evidence while retaining its history (Zhou et al., 2026b), and TOKI formalizes write-time conflict resolution with provenance (Wang, 2026). Our focus is the shared source, scope, and lifecycle checks that govern answer-context inclusion across retrieval and state-read paths.

Retrieval and compression. Retrieval-augmented generation conditions answers on selected passages (Lewis et al., 2020). ReadAgent stores compressed gist memories and returns to original passages when a task requires details (Lee et al., 2024). HippoRAG 2 integrates passages with graph-based retrieval to support factual and associative memory (Gutiérrez et al., 2025b). LightMem combines compression, topic organization, and offline consolidation to reduce online memory overhead (Fang et al., 2026). SimpleMem uses structured compression and intent-aware retrieval

planning to construct compact question-specific context (Liu et al., 2026b); Omni-SimpleMem explores memory designs through an automated research procedure (Liu et al., 2026a). EVStateMem uses summaries as retrieval hints and resolves candidates to attributed source text. Complementary query branches improve coverage within shared scope constraints and a final evidence budget. State-derived and generally retrieved candidates then undergo the same source and lifecycle checks before generation.

Evaluation protocols. MemOps evaluates lifecycle operations through traces of their triggers, targets, scopes, state transitions, and supporting evidence (Hao et al., 2026), complementing final-answer evaluation with operation-level diagnosis. LoCoMo evaluates question answering over long conversations, including questions that require combining events or resolving temporal relationships (Maharana et al., 2024). LongMemEval explicitly includes cross-session inference, temporal reasoning, knowledge updates, and abstention, making it relevant to changing conversational information (Wu et al., 2025). MemEval documents sensitivity to answer models, retrieval settings, and scoring choices (ProsusAI, 2026). These protocols assess whether a system can recover and interpret the information needed for an answer. TimlMemBench complements them with targeted judgments of writing, revocation, replacement, preservation, scope, provenance, boundaries, activation, and restraint against unsolicited use (Zhou et al., 2026a). Its behavior tests ask whether a response follows the applicable memory-handling instruction, including when relevant information should be withheld. We therefore use both types of evidence: conversational question answering measures recovery and reasoning, while the behavior axes examine adherence to update and use constraints.

6 LIMITATIONS

Structured-state coverage. The core state path uses a fixed set of relation definitions. Its coverage depends on the available definitions and on resolving the subject, scope, and time of an update. Information outside this coverage remains accessible through source-linked memory retrieval. Broader structured-state coverage requires extending these definitions and validating the resulting updates; the experimental vocabulary-extension module in Appendix A.2 addresses this direction.

Evaluation coverage. LoCoMo, LongMemEval, and TimlMemBench rely primarily on synthetic conversations or constructed scenarios and do not fully represent natural interactions with real users. Their limited scale and temporal span also prevent a comprehensive assessment of information accumulation, state changes, and memory maintenance over sustained long-term interaction. This is a shared limitation of existing evaluations of long-term memory.

Privacy protection. Withdrawal, scope constraints, and evidence checks govern which memories are available for answering. These controls do not guarantee that generated answers will be free of sensitive information. Privacy protection requires continued testing of both stored information and generated responses, including unintended disclosure under restrictive use policies. Retention and deletion policies must cover source conversations as well as derived memories; response-level behavior scores do not establish that stored information has been physically deleted.

7 CONCLUSION

Long-term conversational memory requires not only retrieving relevant information but also determining whether it remains usable at the current time and within the authorized scope. To this end, EVStateMem links source records, versioned states, Current View, and answer-time evidence checks to govern memory updates and use. LoCoMo and LongMemEval assess recovery and interpretation of conversational evidence; TimlMemBench tests responses to update and use instructions. Results show strengths in single-hop recall, preference and knowledge-update answering, and selective memory use, with remaining weaknesses in writing and recovering provenance.

Future work will explore non-generative decision models that return choices or scores for memory updates, candidate verification, and use decisions. The aim is to reduce latency and cost while retaining source checks, versioned states, and scope constraints.

AI USE STATEMENT

GenAI tools were used during manuscript preparation to improve wording, spelling, grammar, and clarity. The authors are responsible for the paper's intellectual contributions and final content.

ETHICS STATEMENT

Our evaluation uses the publicly available LoCoMo, LongMemEval, and TimlMemBench benchmarks. This study evaluates memory systems on existing benchmark data without collecting new private conversations. Conversational memory raises privacy concerns about retaining and using personal information, motivating our focus on withdrawal, scope constraints, and selective use.

REPRODUCIBILITY STATEMENT

Section 3 describes the memory formation, state update, and evidence selection procedures. Section 4 and Appendix A document the evaluation datasets, model configurations, retrieval settings, and scoring protocols. The corresponding table notes and appendix describe the sources of baseline results and the configurations used for local reruns. The system implementation and usage instructions are available at `https://anonymous.4open.science/r/anonymous_ICLR-E7C5`.

REFERENCES

AgentScope Contributors. ReMe: Memory management kit for agents. Open-source software, 2026. URL https://github.com/agentscope-ai/ReMe. Accessed September 10, 2026.

Seungju Back, Dongwoo Lee, Naun Kang, Taehee Lee, S. K. Hong, Youngjune Gwon, and Sungjin Ahn. Understanding LoRA as Knowledge Memory: An Empirical Analysis. *arXiv preprint arXiv:2603.01097*, 2026. URL https://arxiv.org/abs/2603.01097.

Zijie Cao, Xijun Qu, Zhicheng Gu, Xiaoshu Chen, Duanyang Yuan, Yanning Hou, Sihang Zhou, Jianxing Gong, Jian Huang, and Yang Mei. AutoViewMem: Self-Configuring Orthogonal Views for Conversational Long-Term Memory. *arXiv preprint arXiv:2609.21940*, 2026a. URL https://arxiv.org/abs/2609.21940.

Zouying Cao, Jiaji Deng, Li Yu, Weikang Zhou, Zhaoyang Liu, Bolin Ding, and Hai Zhao. Remember me, refine me: A dynamic procedural memory framework for experience-driven agent evolution. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16803–16822. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.findings-acl.829. URL https://aclanthology.org/2026.findings-acl.829/.

Han Chen, Zining Zhang, Wenqi Pei, Bingsheng He, Ming Wu, Jason Zeng, Michael Heinrich, Wei Wu, and Hongbao Zhang. MemForest: An efficient agent memory system with hierarchical temporal indexing. *arXiv preprint arXiv:2605.23986*, 2026. URL https://arxiv.org/abs/2605.23986v3.

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

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. URL https://arxiv.org/abs/2504.19413.

Claude-Mem Contributors. Claude-Mem. Open-source software, 2025. URL https://github.com/thedotmack/claude-mem. Accessed September 24, 2026.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and efficient memory-augmented generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.18866.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. *arXiv preprint arXiv:2405.14831*, 2025a. URL https://arxiv.org/abs/2405.14831.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025b. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Xixuan Hao, Zeyu Zhang, Zehao Lin, Yihang Sun, Ziliang Guo, Xichong Zhang, Yuxuan Liang, Feiyu Xiong, and Zhiyu Li. MemOps: Benchmarking lifecycle memory operations in long-horizon conversations. *arXiv preprint arXiv:2607.12893*, 2026. URL https://arxiv.org/abs/2607.12893.

Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, and Yafeng Deng. EverMemOS: A Self-Organizing Memory Operating System for Structured Long-Horizon Reasoning. *arXiv preprint arXiv:2601.02163*, 2026a. URL https://arxiv.org/abs/2601.02163.

Yuyang Hu, Jiongnan Liu, Jiejun Tan, Yutao Zhu, and Zhicheng Dou. Memory Matters More: Event-Centric Memory as a Logic Map for Agent Searching and Reasoning. *arXiv preprint arXiv:2601.04726*, 2026b. URL https://arxiv.org/abs/2601.04726.

Zhanghao Hu, Qinglin Zhu, Runcong Zhao, Di Liang, Hanqi Yan, Yulan He, and Lin Gui. Beyond RAG for Agent Memory: Retrieval by Decoupling and Aggregation. *arXiv preprint arXiv:2602.02007*, 2026c. URL https://arxiv.org/abs/2602.02007.

Weiquan Huang, Zixuan Wang, Hehai Lin, Sudong Wang, Bo Xu, Qian Li, Beier Zhu, Linyi Yang, and Chengwei Qin. AMA: Adaptive Memory via Multi-Agent Collaboration. *arXiv preprint arXiv:2601.20352*, 2026. URL https://arxiv.org/abs/2601.20352.

Resham Joshi. Eywa: Provenance-grounded long-term memory for AI agents. *arXiv preprint arXiv:2605.30771*, 2026. URL https://arxiv.org/abs/2605.30771.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

LangChain. LangMem. Open-source software, 2025. URL https://github.com/langchain-ai/langmem. Accessed September 10, 2026.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. A human-inspired reading agent with gist memory of very long contexts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26396–26415. PMLR, 2024. URL https://proceedings.mlr.press/v235/lee24c.html.

Jingdi Lei, Di Zhang, Junxian Li, Weida Wang, Kaixuan Fan, Xiang Liu, Qihan Liu, Xiaoteng Ma, Baian Chen, and Soujanya Poria. δ-mem: Efficient Online Memory for Large Language Models. *arXiv preprint arXiv:2605.12357*, 2026. URL https://arxiv.org/abs/2605.12357.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Junfeng Liao, Qizhou Wang, Jianing Zhu, Bo Du, Rui Yan, and Xiuying Chen. Belief memory: Agent memory under partial observability. *arXiv preprint arXiv:2605.05583*, 2026. URL https://arxiv.org/abs/2605.05583.

Minhua Lin, Zhiwei Zhang, Hanqing Lu, Hui Liu, Xianfeng Tang, Qi He, Xiang Zhang, and Suhang Wang. MemMA: Coordinating the Memory Cycle through Multi-Agent Reasoning and In-Situ Self-Evolution. *arXiv preprint arXiv:2603.18718*, 2026. URL https://arxiv.org/abs/2603.18718.

Jiaqi Liu, Zipeng Ling, Shi Qiu, Yanqing Liu, Siwei Han, Peng Xia, Haoqin Tu, Zeyu Zheng, Cihang Xie, Charles Fleming, Mingyu Ding, and Huaxiu Yao. Omni-SimpleMem: Autoresearch-guided discovery of lifelong multimodal agent memory. *arXiv preprint arXiv:2604.01007*, 2026a. URL https://arxiv.org/abs/2604.01007.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. SimpleMem: Efficient lifelong memory for LLM agents. *arXiv preprint arXiv:2601.02553*, 2026b. URL https://arxiv.org/abs/2601.02553.

Junming Liu, Yifei Sun, Weihua Cheng, Haodong Lei, Yirong Chen, Licheng Wen, Xuemeng Yang, Daocheng Fu, Pinlong Cai, Nianchen Deng, Yi Yu, Shuyue Hu, Botian Shi, and Ding Wang. MemVerse: Multimodal Memory for Lifelong Learning Agents. *arXiv preprint arXiv:2512.03627*, 2026c. URL https://arxiv.org/abs/2512.03627.

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-Memory: Recalling and Post-thinking Enable LLMs with Long-Term Memory. *arXiv preprint arXiv:2311.08719*, 2023. URL https://arxiv.org/abs/2311.08719.

Wenquan Ma, Jiayan Nan, Wenlong Wu, and Yize Chen. What Deserves Memory: Adaptive Memory Distillation for LLM Agents. *arXiv preprint arXiv:2508.03341*, 2026. URL https://arxiv.org/abs/2508.03341.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

MemPalace Contributors. MemPalace. Open-source software, 2026. URL https://github.com/MemPalace/mempalace. Accessed September 10, 2026.

NevaMind AI. memU. Open-source software, 2026. URL https://github.com/NevaMind-AI/memU. Accessed September 14, 2026.

OceanBase. PowerMem. Open-source software, 2026. URL https://github.com/oceanbase/powermem. Repository now redirects to PowerContext; accessed September 10, 2026.

OpenClaw Contributors. OpenClaw. Open-source software, 2026. URL https://github.com/openclaw/openclaw. Accessed September 24, 2026.

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

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression. *arXiv preprint arXiv:2403.12968*, 2024. URL https://arxiv.org/abs/2403.12968.

Daivik Patel and Shrenik Patel. ENGRAM: Effective, Lightweight Memory Orchestration for Conversational Agents. *arXiv preprint arXiv:2511.12960*, 2026. URL https://arxiv.org/abs/2511.12960.

ProsusAI. MemEval: Fair evaluation framework for agent memory. Open-source software, 2026. URL https://github.com/ProsusAI/MemEval. Accessed September 10, 2026.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025. URL https://arxiv.org/abs/2501.13956.

Zitong Shi, Yixuan Tang, and Anthony Kum Hoe Tung. A-TMA: Decoupling state-aware memory failures in long-term agent memory. *arXiv preprint arXiv:2607.01935*, 2026. URL https://arxiv.org/abs/2607.01935.

Jingwei Sun, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. Rethinking how to remember: Beyond atomic facts in lifelong LLM agent memory. *arXiv preprint arXiv:2605.19952*, 2026. URL https://arxiv.org/abs/2605.19952.

Kaixiang Wang, Yidan Lin, Jiong Lou, Zhaojiacheng Zhou, Bunyod Suvonov, and Jie Li. E-mem: Multi-agent based Episodic Context Reconstruction for LLM Agent Memory. *arXiv preprint arXiv:2601.21714*, 2026a. URL https://arxiv.org/abs/2601.21714.

Shu Wang, Edwin Yu, Oscar Love, Tom Zhang, Tom Wong, Steve Scargall, and Charles Fan. MemMachine: A Ground-Truth-Preserving Memory System for Personalized AI Agents. *arXiv preprint arXiv:2604.04853*, 2026b. URL https://arxiv.org/abs/2604.04853.

Xinyuan Wang, Wenyu Mao, Junkang Wu, Xiang Wang, and Xiangnan He. R^2-Mem: Reflective Experience for Memory Search. *arXiv preprint arXiv:2605.13486*, 2026c. URL https://arxiv.org/abs/2605.13486.

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

Rubin Wei, Jiaqi Cao, Jiarui Wang, Jushi Kai, Qipeng Guo, Bowen Zhou, and Zhouhan Lin. MLP Memory: A Retriever-Pretrained Memory for Large Language Models. *arXiv preprint arXiv:2508.01832*, 2026. URL https://arxiv.org/abs/2508.01832.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, pp. 86809–86836, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d813d324dbf0598bbdc9c8e79740ed01-Paper-Conference.pdf.

Menglin Xia, Xuchao Zhang, Shantanu Dixit, Paramaguru Harimurugan, Rujia Wang, Victor Ruhle, Robert Sim, Chetan Bansal, and Saravan Rajmohan. Memora: A Harmonic Memory Representation Balancing Abstraction and Specificity. *arXiv preprint arXiv:2602.03315*, 2026. URL https://arxiv.org/abs/2602.03315.

Yilin Xiao, Zhehan Zhu, Yujing Zhang, Jin Chen, Zijin Hong, Luyao Zhuang, Qinggang Zhang, Shengyuan Chen, Xiaocao Ouyang, Lingfei Ren, and Xiao Huang. Zero-Mem: Zero-Token Memory Operations for LLM Agents. *arXiv preprint arXiv:2607.29377*, 2026. URL https://arxiv.org/abs/2607.29377.

Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, wenlin zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, and Tong Xu. From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents. *arXiv preprint arXiv:2505.19549*, 2025a. URL https://arxiv.org/abs/2505.19549.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17577–17604, 2025b. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Meriem Yacoubi, Pia Schmidt, Nenad Petrovic, Ahmed Frikha, Martin Kirchhoff, and Alois Knoll. MemoryLACE: Memory lifecycle-aware consolidation and evidence retrieval. *arXiv preprint arXiv:2609.03201*, 2026. URL https://arxiv.org/abs/2609.03201.

B. Y. Yan, Chaofan Li, Hongjin Qian, Shuqi Lu, and Zheng Liu. General Agentic Memory Via Deep Research. *arXiv preprint arXiv:2511.18423*, 2025. URL https://arxiv.org/abs/2511.18423.

Shannan Yan, Jingchen Ni, Leqi Zheng, Jiajun Zhang, Peixi Wu, Dacheng Yin, Jing Lyu, Chun Yuan, and Fengyun Rao. AdaMem: Adaptive User-Centric Memory for Long-Horizon Dialogue Agents. *arXiv preprint arXiv:2603.16496*, 2026a. URL https://arxiv.org/abs/2603.16496.

Sikuan Yan, Ahmed Bahloul, Ercong Nie, Susanna Schwarzmann, Riccardo Trivisonno, Volker Tresp, and Yunpu Ma. Memory-R2: Fair Credit Assignment for Long-Horizon Memory-Augmented LLM Agents. *arXiv preprint arXiv:2605.21768*, 2026b. URL https://arxiv.org/abs/2605.21768.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning. *arXiv preprint arXiv:2508.19828*, 2026c. URL https://arxiv.org/abs/2508.19828.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. MemAgent: Reshaping Long-Context LLM with Multi-Conv RL-based Memory Agent. *arXiv preprint arXiv:2507.02259*, 2026. URL https://arxiv.org/abs/2507.02259.

Yanwei Yue, Boci Peng, Xuanbo Fan, Jiaxin Guo, Qiankun Li, and Yan Zhang. Mem-T: Densifying Rewards for Long-Horizon Memory Agents. *arXiv preprint arXiv:2601.23014*, 2026. URL https://arxiv.org/abs/2601.23014.

Guibin Zhang, Muxin Fu, and Shuicheng Yan. MemGen: Weaving Generative Latent Memory for Self-Evolving Agents. *arXiv preprint arXiv:2509.24704*, 2025. URL https://arxiv.org/abs/2509.24704.

Jiaquan Zhang, Chaoning Zhang, Shuxu Chen, Zhenzhen Huang, Pengcheng Zheng, Zhicheng Wang, Ping Guo, Fan Mo, Sung-Ho Bae, Jie Zou, Jiwei Wei, and Yang Yang. Lightweight LLM agent memory with small language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12914–12929. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.588. URL https://aclanthology.org/2026.acl-long.588/.

Lei Zheng, Weinan Song, Daili Li, and Yanming Yang. To Know is to Construct: Schema-Constrained Generation for Agent Memory. *arXiv preprint arXiv:2604.20117*, 2026. URL https://arxiv.org/abs/2604.20117.

Yicong Zheng, Kevin L. McKee, Thomas Miconi, Zacharie Bugaud, Mick van Gelderen, and Jed McCaleb. Goal-directed search outperforms goal-agnostic memory compression in long-context memory tasks. *arXiv preprint arXiv:2511.21726*, 2025. URL https://arxiv.org/abs/2511.21726.

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

Yijie Zhong, Yunfan Gao, and Haofen Wang. HingeMem: Boundary guided long-term memory with query adaptive retrieval for scalable dialogues. In *Proceedings of the ACM Web Conference 2026*, 2026. doi: 10.1145/3774904.3792089. URL https://arxiv.org/abs/2604.06845.

Fangxing Zhou, Yifei Sun, Jiachen Li, Jinyu Xu, Yuhan Wang, Ruixin Li, and Qing Xie. TimlMem-Bench: A diagnostic benchmark for governed use of long-term memory in dialogue agents. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026a. URL https://github.com/lyirs/TimlMemBench. To appear.

Yan Zhou, Yue Ouyang, Kaiyang Zheng, and Suncheng Xiang. TEPA: Revoking stale memories for conflict-robust language agents. *arXiv preprint arXiv:2608.07429*, 2026b. URL https://arxiv.org/abs/2608.07429.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. MEM1: Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. *arXiv preprint arXiv:2506.15841*, 2025. URL https://arxiv.org/abs/2506.15841.

Jinchang Zhu, Jindong Li, Cheng Zhang, Jiahong Liu, and Menglin Yang. HeLa-Mem: Hebbian learning and associative memory for LLM agents. *arXiv preprint arXiv:2604.16839*, 2026. URL https://arxiv.org/abs/2604.16839.

A ADDITIONAL EVALUATION DETAILS

A.1 LONGMEMEVAL CONFIGURATIONS

Configurations marked † apply length corrections to individual outputs exceeding the answer-length limit; parsing rules and the judging model remain unchanged.

S-cleaned contains 70 single-session user, 56 assistant, 30 preference, 133 multi-session, 133 temporal, and 78 knowledge-update questions. Overall accuracy pools all 500 questions; the 30 abstention questions overlap these categories. EVStateMem ingests raw conversations, with temperature 0 and output limits of 4,096 reader tokens and 10 judge tokens.

BM25 and RAG+BGE-M3 retrieve 20 source turns; BM25 uses $k_1 = 1.5$ and $b = 0.75$. MemPalace uses MiniLM and all-role hybrid retrieval with a limit of 20. This configuration evaluates question answering over both speakers' histories.

MemoryBank-summary uses MemoryBank's session and overall summarization prompts (Zhong et al., 2024), with BGE-M3 cosine retrieval over summaries and a Top-20 limit. Personality modeling, forgetting, and raw-turn retrieval are excluded. Summary generation uses Terra at temperature zero with a 1,024-token output limit; answering and judging use the shared Terra protocol. All 500 questions are evaluated.

The Qwen configuration uses Qwen3-4B-Instruct-2507 (Q8, llama.cpp) to read a fixed pool of retrieved source blocks. Evidence selection operates on groups of four blocks, preserves the original top four, and adds a source-time index. The reader drafts an answer and reviews it against the selected evidence for every question.

We use the public judge prompt adopted by MemForest (Mem0 commit `7ba1bd3`), with the judging temperature set to zero and thinking disabled. On the same fixed answers, the original task-specific rubric and this public prompt yield accuracies of 66.00% and 71.40%, respectively. EverMemOS is described by Hu et al. (2026a); we use its results reported in MemForest. MemForest evaluates the full cleaned dataset, using FP8 readers and the respective native memory pipelines, without reporting baseline abstention scores. Thus, although the comparison uses the same reader model family and public judge prompt, model precision and memory pipelines still differ across systems (Chen et al., 2026).

Table 5: LongMemEval S-cleaned accuracy (%). SS-U/A/P: single-session user/assistant/preference; Multi: multi-session; Temp: temporal; Update: knowledge update. Bold and underline mark the best and second-best distinct scores within each model group. †: answer-length interventions (Appendix A.1).

Method	Overall	SS-U	SS-A	SS-P	Multi	Temp	Update
Answer: **GPT-5.6-terra**							Judge: GPT-5.6-terra
BM25†	71.00	87.14	85.71	26.67	60.90	70.68	80.77
RAG+BGE-M3†	80.20	<u>88.57</u>	89.29	33.33	<u>80.45</u>	<u>78.95</u>	<u>85.90</u>
MemPalace†	71.80	85.71	<u>91.07</u>	30.00	62.41	68.42	83.33
MemoryBank-summary†	58.40	78.57	58.93	30.00	60.15	45.11	70.51
LightMem	67.60	85.71	19.64	30.00	72.18	72.93	83.33
Full context†	**84.00**	**94.29**	**96.43**	<u>36.67</u>	**81.95**	**86.47**	83.33
EVStateMem	<u>81.60</u>	82.86	<u>91.07</u>	**73.33**	76.69	<u>78.95</u>	**89.74**
Answer: **Qwen3-4B-Instruct-2507**							Judge: DeepSeek-V4-Flash
MemForest-Planner	**72.60**	**95.71**	83.93	<u>70.00</u>	59.40	**70.68**	70.51
MemForest-Embed	69.40	88.57	83.93	53.33	60.90	**70.68**	60.26
EverMemOS	61.20	84.29	69.64	66.67	55.64	48.12	64.10
LightMem	68.80	<u>94.29</u>	35.71	**76.67**	**65.41**	64.66	**79.49**
MemoryOS	64.20	87.14	<u>89.29</u>	46.67	57.14	60.90	50.00
MemPalace	60.00	91.43	33.93	36.67	<u>63.91</u>	52.63	65.38
Mem0	44.80	82.86	26.79	33.33	42.11	38.35	43.59
Zep Local	69.60	<u>94.29</u>	**98.21**	53.33	59.40	54.89	<u>75.64</u>
EVStateMem	<u>71.40</u>	92.86	**98.21**	53.33	51.13	<u>68.42</u>	**79.49**

A.2 Experimental relation-vocabulary extension

The experimental extension introduces new relation definitions when the existing vocabulary does not cover a source-supported assertion. An extraction step proposes assertions and candidate definitions; a verification step checks source support, persistence, dates, and update targets. Semantically equivalent definitions reuse existing identifiers and aliases.

Candidate definitions must satisfy structural, type, scope, and temporal constraints, with quoted evidence traceable to a linked source. Accepted definitions describe reusable relations; their values remain in scope-bound state records. Vocabulary expansion preserves the semantics of assert, supersede, and retract.

Extraction and verification use the same model through a dedicated provider configuration. Query matching uses relation identifiers and aliases. The two-stage procedure adds inference latency and token consumption.

The system owner can review source-linked proposals before admitting definitions to the active catalog. Approved definitions become available to the runtime vocabulary and Current View materialization through explicit activation. Each decision is recorded, and catalog admission preserves the existing Current View.

A.3 LoCoMo configurations

Both EVStateMem runs ingest raw conversations and retrieve up to 20 primary items. Each answer model also serves as judge.

The 446 adversarial questions use free-response abstention F1: 76.91 for Terra and 37.89 for GPT-4o. They are reported separately because several baselines use multiple-choice questions. Tables 6 and 7 cover the four non-adversarial categories.

Terra baselines use published methods or released software: MemEval for PropMem/OpenClaw (ProsusAI, 2026), Graphiti/Zep (Rasmussen et al., 2025), LangMem (LangChain, 2025), MemPalace (MemPalace Contributors, 2026), PowerMem (OceanBase, 2026), ReMe (AgentScope Contributors, 2026), and memU (NevaMind AI, 2026). The marked adaptations use local MiniLM for LangMem/SimpleMem, BGE-M3 for ReMe/memU (Chen et al., 2024), and only MemoryBank's summary component (Zhong et al., 2024). RAG+BGE-M3 retrieves conversation turns.

Primary retrieval limits are 10 for A-MEM, Graphiti, LangMem, and MemoryOS; Mem0, MemPalace, OpenClaw, PowerMem, ReMe, MemoryBank-summary, and RAG+BGE-M3 use 20. LightMem uses 60 combined items; PropMem uses 30 propositions and 3 chunks. SimpleMem uses 25/5/5 semantic/keyword/structured candidates; memU uses 5/5/5 category/item/resource limits. These limits count different retrieval units.

GPT-4o score sources are identified below Table 7; superscripts also cover each source's baseline reproductions. ReadAgent and HippoRAG2 follow Lee et al. (2024) and Gutiérrez et al. (2025b). Among HingeMem-source rows, only RAG Top-20 uses category-specific answer-format prompts; Mem0_g is the graph-memory variant. Missing category metrics remain –.

Omni-SimpleMem (Liu et al., 2026a) uses local MiniLM and retrieval limits of 20/20/20/30 for single-hop/multi-hop/open-domain/temporal questions. Its local GPT-4o run covers all 1,540 questions in these categories, with GPT-4o judging.

Table 6: Complete four-category LoCoMo metrics with GPT-5.6-terra (%): F1, B1, and semantic Judge accuracy J. All rows are local runs. Bold and underline mark the best and second-best distinct scores per column; ties share the same mark; † identifies the adaptations described above.

		Single-hop			Multi-hop			Open-domain			Temporal		
Method	Overall F1	F1	B1	J	F1	B1	J	F1	B1	J	F1	B1	J
A-MEM	45.00	49.77	43.47	65.40	33.66	26.95	33.69	26.47	22.94	46.88	48.01	42.14	55.76
Graphiti	41.29	45.61	39.60	49.58	29.52	24.28	24.47	29.63	24.08	41.67	43.82	38.42	29.91
LangMem†	14.51	15.83	12.85	27.82	14.20	11.54	16.31	12.50	11.07	23.96	11.93	10.43	16.51
LightMem	56.36	58.17	50.32	76.10	42.40	34.96	43.97	33.74	27.95	50.00	**70.67**	**64.66**	78.19
Mem0	24.96	30.43	21.67	84.30	26.19	18.98	**82.27**	18.45	14.34	58.33	11.50	8.32	20.56
MemoryBank-summary†	45.41	45.54	38.01	61.71	40.86	32.58	41.84	27.67	23.30	41.67	54.37	47.25	37.69
MemoryOS	48.85	52.86	46.18	66.23	38.97	32.12	39.72	32.81	28.20	50.00	51.80	45.17	44.86
MemPalace	49.86	54.57	47.76	66.11	34.17	27.98	33.69	32.70	27.31	53.13	56.42	52.12	65.42
memU†	33.11	34.70	29.28	47.09	30.92	24.89	26.95	22.18	18.74	37.50	34.15	29.52	32.40
Omni-SimpleMem	51.69	54.47	46.70	76.58	44.35	37.22	48.23	**34.05**	**29.07**	47.92	56.13	49.33	78.50
OpenClaw	45.93	53.67	46.27	64.21	38.31	31.03	30.85	27.27	21.90	43.75	37.93	32.51	39.56
PowerMem	48.99	51.64	44.75	62.90	32.73	26.63	31.91	30.39	24.93	45.83	61.91	54.91	71.96
PropMem	56.17	61.18	53.99	73.48	39.34	32.28	35.11	32.77	27.07	52.08	64.85	59.11	74.45
RAG+BGE-M3	52.79	58.52	51.06	73.13	34.98	28.51	34.04	27.67	22.55	46.88	60.95	55.83	68.85
ReMe†	21.71	21.34	12.68	**87.51**	24.16	15.81	62.77	9.33	5.51	**63.54**	24.24	15.18	73.52
SimpleMem†	53.03	55.22	47.46	83.35	43.06	34.94	54.96	27.61	24.39	50.00	63.67	57.62	**79.75**
EVStateMem	**61.34**	**69.11**	**62.13**	83.12	**52.30**	**45.57**	43.62	30.03	25.48	43.75	58.32	53.01	70.40

Table 7: Full available four-category LoCoMo metrics for representative GPT-4o configurations (%). – denotes an unreported category metric. Superscripts identify the score source, including that source's baseline reproductions. Among the g rows, only RAG Top-20 uses category-specific answer-format prompts. Judge settings differ across reports. Bold and underline mark the best and second-best distinct scores per column; ties share the same mark.

		Single-hop			Multi-hop			Open-domain			Temporal		
Method	Overall F1	F1	B1	J	F1	B1	J	F1	B1	J	F1	B1	J
A-MEM^{a}	41.75	48.43	42.97	–	32.86	23.76	–	17.10	15.84	–	39.41	31.23	–
BeliefMem^{b}	–	51.81	43.80	–	41.24	35.68	–	27.60	**23.78**	–	50.81	**45.05**	–
HeLa-Mem^{h}	46.41	49.69	44.08	–	39.12	29.82	–	24.38	19.24	–	50.79	44.54	–
HingeMem^{g}	**57.06**	61.10	–	**78.80**	**53.60**	–	**62.80**	**30.70**	–	46.40	57.40	–	**66.90**
HippoRAG2^{g}	49.22	54.40	–	78.50	35.40	–	52.80	23.40	–	35.20	55.50	–	61.00
LangMem^{g}	50.41	54.70	–	70.10	47.40	–	58.50	27.80	–	38.30	48.60	–	51.00
LightMem^{t}	47.37	46.60	36.72	–	45.16	35.30	–	23.20	10.56	–	**58.57**	43.60	–
LoCoMo^{a}	41.67	61.56	54.19	–	28.00	18.47	–	16.47	14.80	–	9.09	5.78	–
Mem0^{t}	37.74	39.84	30.02	–	32.36	25.52	–	18.50	14.50	–	42.70	32.48	–
MemGPT^{a}	42.78	60.16	53.35	–	30.36	22.83	–	12.24	11.87	–	17.29	13.18	–
MemoryBank^{a}	6.63	8.28	7.10	–	6.49	4.69	–	6.43	5.30	–	2.47	2.43	–
MemoryOS^{t}	40.60	43.67	38.26	–	31.76	22.52	–	18.06	12.91	–	47.08	38.31	–
Naive RAG^{t}	25.88	35.03	26.67	–	15.35	14.36	–	9.09	9.03	–	16.17	11.48	–
Omni-SimpleMem^{r}	47.28	51.74	44.07	69.08	38.67	32.02	44.33	25.19	20.90	39.58	49.78	42.78	56.39
ReadAgent^{a}	10.90	12.46	10.29	–	14.61	9.95	–	8.84	8.37	–	4.16	3.19	–
SimpleMem^{s}	44.33	45.41	39.25	–	35.89	32.83	–	18.23	16.34	–	56.71	20.57	–
Zep^{g}	20.21	20.60	–	70.00	22.00	–	52.40	6.60	–	40.40	21.70	–	52.60
Mem0 (HingeMem)^{g}	44.53	45.10	–	56.70	42.70	–	48.20	27.70	–	**47.20**	49.70	–	50.10
Mem0$_g$^{g}	43.89	45.30	–	55.10	40.40	–	48.50	28.40	–	41.00	47.90	–	52.00
RAG Top-20^{g}	34.60	37.70	–	52.50	34.90	–	48.90	21.90	–	41.60	30.00	–	22.70
EVStateMem	54.01	**63.16**	**55.03**	77.05	45.71	**38.77**	49.65	27.29	23.49	32.29	45.32	39.81	48.60

Sources: a A-MEM (Xu et al., 2025b); b BeliefMem (Liao et al., 2026); h HeLa-Mem (Zhu et al., 2026); s SimpleMem (Liu et al., 2026b); t TriMem (Sun et al., 2026); g HingeMem (Zhong et al., 2026); r local rerun described above.

A.3.1 CROSS-MODEL OVERVIEW OF REPORTED RESULTS

Figure 2 summarizes LoCoMo category F1 for 49 methods or configurations across 23 model groups, covering 224 reported method–model pairs. Model groups have at least six available method labels; a method may have results for just one model. Generic controls, undisclosed answer models, Judge-only records, and labels without verified year metadata are excluded.

When several reports cover the same method and model, one complete record is selected by source priority: local run, the method's own report, then alphabetical source order. Selection is independent of scores. Evaluation protocols vary across reports, so the figure presents coverage and score distributions.

Additional method references. The overview also includes the following methods, cited here separately from their score sources: TiM (Liu et al., 2023); HippoRAG (Gutiérrez et al., 2025a); LLMLingua-2 (Pan et al., 2024); ENGRAM (Patel & Patel, 2026); GAM (Yan et al., 2025); MEM1 (Zhou et al., 2025); MemAgent (Yu et al., 2026); MemGAS (Xu et al., 2025a); MemGen (Zhang et al., 2025); MemMachine (Wang et al., 2026b); MemVerse (Liu et al., 2026c); Memory-R1 (Yan et al., 2026c); Nemori (Ma et al., 2026).

AMA (Huang et al., 2026); AdaMem (Yan et al., 2026a); AutoViewMem (Cao et al., 2026a); CompassMem (Hu et al., 2026b); Context2LoRA (Back et al., 2026); E-mem (Wang et al., 2026a); MemMA (MEMMA LM) (Lin et al., 2026); MLP Memory (Wei et al., 2026); Memora (Xia et al., 2026); Memory-R2 (Yan et al., 2026b); R^2-Mem (Wang et al., 2026c); SCG-Mem (Zheng et al., 2026); Zero-Mem (Xiao et al., 2026); xMemory (Hu et al., 2026c).

Software entries follow their official projects: Claude-Mem (Claude-Mem Contributors, 2025); OpenClaw (OpenClaw Contributors, 2026). PropMem and the evaluated OpenClaw configuration follow MemEval (ProsusAI, 2026).

Additional reproduced scores are drawn from Mem-T (Yue et al., 2026), δ-Mem (Lei et al., 2026), and SUMER (Zheng et al., 2025), alongside the cited method reports. A reproduced baseline retains its original method identity while its scores are attributed to the reporting paper.

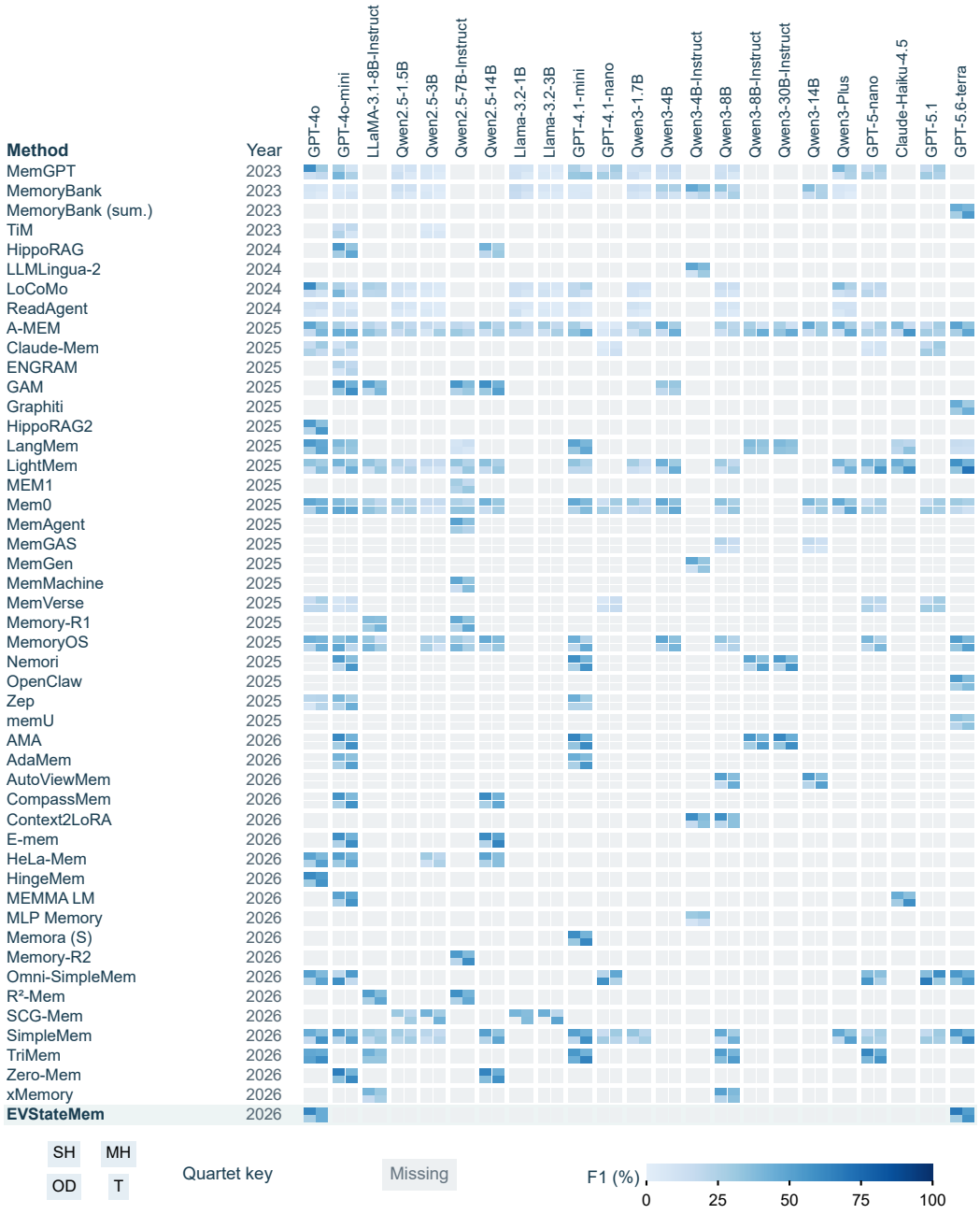

Figure 2: Cross-model LoCoMo category F1 (%). Each quartet places single-hop (SH) and multi-hop (MH) above open-domain (OD) and temporal (T) scores. Gray denotes missing scores; all cells share one color scale. Methods follow first public year and models approximate release chronology.

A.4 TIMLMEMBENCH QUESTION-ANSWERING CONTROLS

The canonical-pool control supplies the same sample-local memory items to each system before retrieval and answering. Tables 8 and 9 include all fourteen baseline configurations and EVStateMem, using 979 Chinese and 736 English questions. Baseline configurations follow Zhou et al. (2026a); language-level scores come from its accompanying scorecard.

Table 8: Complete Chinese canonical-pool question-answering results on TimlMemBench. Scores are percentages except token counts and latency. Bold and underline mark the best and second-best distinct percentage scores per column; ties share the same mark.

System	*J*	Ans.	Abs.	Ret.	Comb.	F1	B1	Tok./Q	E2E (s)
No memory	11.1	2.0	85.2	0.8	0.7	13.3	13.3	1552.5	23.9
Full context	81.7	81.6	82.4	**100.0**	81.7	54.2	40.7	4507.6	31.3
Summary	83.9	83.6	86.1	95.5	80.3	56.3	43.6	2985.6	28.3
RAG (BGE-M3)	81.7	81.6	82.4	95.5	79.0	55.8	42.9	3150.8	28.4
MemoryBank (summary)	**97.5**	**97.4**	99.1	95.5	**93.6**	59.8	45.6	3125.9	19.2
Zep	94.1	93.5	99.1	93.8	90.3	55.9	41.5	3811.0	38.0
MemPalace (pool)	72.1	71.0	81.5	78.8	65.2	49.5	37.6	3972.2	36.7
MemPalace (native)	71.1	69.2	86.1	76.3	63.3	47.6	35.9	3916.7	35.8
Mem0	80.5	80.0	84.3	95.9	77.4	54.4	41.2	3420.3	37.2
LangMem	81.4	81.3	82.4	96.7	79.1	54.9	41.7	3270.1	32.0
Graphiti	79.0	78.6	81.5	93.9	75.8	53.3	40.6	3493.6	32.6
A-MEM	79.6	79.1	83.3	95.0	76.3	54.2	41.3	3226.6	31.5
SimpleMem	79.8	79.0	86.1	96.6	77.4	53.9	40.8	3183.3	30.9
PowerMem	80.1	79.4	85.2	96.6	77.8	54.0	41.0	3362.2	32.8
EVStateMem	87.5	86.0	**100.0**	93.0	83.2	**62.2**	**55.7**	1496.4	10.9

J: semantic Judge accuracy; Ans.: Judge accuracy on answerable questions; Abs.: abstention accuracy; Ret.: memory-item retrieval accuracy; Comb.: correct retrieval and answer; B1: BLEU-1. Tok./Q is tokens per question; E2E is mean end-to-end latency, including retrieval, answer generation, and judging.

Table 9: Complete English canonical-pool question-answering results on TimlMemBench. Scores are percentages except token counts and latency. Bold and underline mark the best and second-best distinct percentage scores per column; ties share the same mark.

System	*J*	Ans.	Abs.	Ret.	Comb.	F1	B1	Tok./Q	E2E (s)
No memory	11.1	6.3	**100.0**	0.1	0.1	7.9	7.9	2087.1	23.7
Full context	91.6	91.3	97.4	**100.0**	**91.6**	52.0	41.0	4767.0	34.2
Summary	91.3	90.8	**100.0**	92.4	86.1	50.2	39.3	3400.7	32.3
RAG (BGE-M3)	92.1	91.7	**100.0**	92.4	86.8	50.7	39.5	3570.2	30.1
MemoryBank (summary)	**93.2**	**93.0**	97.4	92.4	88.0	50.3	39.3	3385.1	30.2
Zep	91.0	90.8	94.7	97.1	88.9	51.2	40.3	4003.0	43.5
MemPalace (pool)	89.7	89.4	94.7	97.7	88.3	50.7	40.4	3792.2	34.4
MemPalace (native)	90.9	90.5	97.4	97.8	89.7	50.5	39.4	3846.4	36.0
Mem0	90.6	90.3	97.4	95.8	88.2	50.3	39.4	3906.3	41.6
LangMem	92.1	92.0	94.7	96.7	89.9	51.1	40.4	3925.7	36.6
Graphiti	91.2	91.0	94.7	97.4	89.5	51.3	40.5	3932.7	36.3
A-MEM	90.8	90.5	94.7	96.5	88.7	51.3	40.6	3696.0	34.7
SimpleMem	90.4	90.3	92.1	96.7	88.3	50.4	39.7	3776.1	35.8
PowerMem	91.6	91.4	94.7	96.7	89.4	51.3	40.5	3867.9	34.8
EVStateMem	81.5	80.5	**100.0**	93.2	77.7	**53.1**	**46.7**	1324.6	15.9

J: semantic Judge accuracy; Ans.: Judge accuracy on answerable questions; Abs.: abstention accuracy; Ret.: memory-item retrieval accuracy; Comb.: correct retrieval and answer; B1: BLEU-1. Tok./Q is tokens per question; E2E is mean end-to-end latency, including retrieval, answer generation, and judging.

Our run uses each question's prescribed retrieval limit and gateway-reported GPT-5.4-mini for answering and judging, with temperature 0, `xhigh` reasoning, and a 32,768-token output limit. Tokens count answer-generation and judge input/output; end-to-end latency includes retrieval, answering, and judging.

A.5 TIMLMEMBENCH RESULTS BY LANGUAGE

Baseline identities and variants follow Zhou et al. (2026a). Its LightMem (SLM) and ReMe rows follow Zhang et al. (2026) and Cao et al. (2026b). LoCoMo's LightMem follows Fang et al. (2026).

For axis a and language ℓ, let $c_{\ell,a}$ count accepted cases and $n_{\ell,a}$ count all scored cases. The combined score is $100(c_{\mathrm{ZH},a} + c_{\mathrm{EN},a})/(n_{\mathrm{ZH},a} + n_{\mathrm{EN},a})$. For each published baseline cell, its percentage and denominator uniquely determine an integer accepted-case count that rounds back to the published value. We pool these counts and round to 0.1 percentage points, with ties to even. Tables 10 and 11 preserve every language-specific score; the reproducibility artifact retains the denominators.

Table 10: TimlMemBench Chinese raw-native behavior scores (%). Baselines are from Zhou et al. (2026a). Bold and underline mark the best and second-best distinct scores per column; ties share the same mark.

System	Write	Revoke	Supersede	Preserve	Scope	Provenance	Boundary	Activation	Proactive
Mem0	36.4	46.3	47.3	44.4	46.0	39.1	50.5	50.9	58.2
MemoryOS	**69.1**	67.0	66.2	64.8	61.0	**73.4**	65.3	65.4	70.4
MemoryBank	43.6	64.9	66.2	57.4	56.0	42.2	70.8	71.3	71.4
MemPalace	7.3	20.7	27.0	24.1	24.0	15.6	28.0	21.5	29.6
Graphiti	18.2	23.4	22.1	25.0	28.0	21.9	29.6	35.1	31.6
Zep	20.0	33.0	36.0	45.4	43.0	31.2	43.0	41.7	49.0
LightMem (SLM)	29.1	33.5	32.4	45.4	40.0	34.4	41.5	38.6	53.1
ReMe	27.3	32.4	24.3	45.4	50.0	29.7	43.0	37.1	51.0
PowerMem	34.5	56.4	45.5	48.1	47.0	40.6	53.0	53.7	57.1
A-MEM	50.9	60.1	61.7	58.3	59.0	56.2	63.9	62.3	71.4
SimpleMem	9.1	26.1	18.0	36.1	39.0	6.2	29.6	26.5	43.9
TriMem	30.9	46.3	36.5	59.3	52.0	37.5	51.1	48.0	58.2
EVStateMem	50.9	**73.9**	**68.5**	**69.4**	**69.0**	60.9	**78.5**	**75.7**	**82.7**

Table 11: TimlMemBench English raw-native behavior scores (%). Baselines are from Zhou et al. (2026a). Bold and underline mark the best and second-best distinct scores per column; ties share the same mark.

System	Write	Revoke	Supersede	Preserve	Scope	Provenance	Boundary	Activation	Proactive
Mem0	47.0	49.0	61.3	46.0	48.8	47.0	60.5	57.6	65.9
MemoryOS	**66.4**	69.1	68.8	60.8	63.3	72.0	63.3	68.1	64.9
MemoryBank	55.2	67.1	71.0	65.3	69.9	65.9	69.4	72.6	72.4
MemPalace	41.0	54.2	60.2	44.9	53.5	56.1	55.1	65.2	61.6
Graphiti	61.2	67.9	74.2	64.5	68.4	66.5	66.0	73.8	74.1
Zep	32.8	42.6	31.2	50.2	52.3	34.8	51.0	53.1	63.8
LightMem (SLM)	44.0	53.4	49.5	55.5	60.5	52.4	58.5	59.0	60.5
ReMe	40.3	47.0	34.4	50.2	55.5	36.0	67.3	52.4	64.3
PowerMem	36.6	53.4	43.0	55.5	59.0	47.0	59.2	65.2	75.1
A-MEM	63.4	71.5	**80.6**	65.3	69.1	**72.6**	**74.1**	73.3	69.7
SimpleMem	46.3	56.2	52.7	55.1	60.5	56.7	63.3	64.8	74.6
TriMem	45.5	59.8	54.8	52.8	55.5	55.5	59.2	61.9	68.6
EVStateMem	53.0	**77.9**	75.3	**77.0**	**79.7**	63.4	73.5	**80.5**	**88.6**

A.6 Sensitivity to Judge prompts and models

Judge prompts. We compare the official LongMemEval task-specific templates with the unified Mem0 `JUDGE_PROMPT` used by MemForest on frozen Qwen answers (Wu et al., 2025; Chen et al., 2026). Within each set, the questions, references or rubrics, answers, and Judge settings are unchanged: DeepSeek-V4-Flash, temperature 0, at most 10 output tokens, thinking disabled, and `yes/no` output. Table 12 shows the sensitivity of fixed answers to the judging prompt.

Table 12: Full LongMemEval accuracy under two judging prompts (%). Answers are fixed within each configuration; Overall includes all 500 questions, with failed answers counted as incorrect. Abs.* is the abstention subset, which overlaps the task categories.

Judge prompt	Overall	SS-U	SS-A	SS-P	Multi	Temp	Update	Abs.*
Aligned reader, strict string parsing								
Official task-specific	57.40	80.00	92.86	43.33	30.83	58.65	60.26	63.33
Mem0 unified	63.20	88.57	92.86	60.00	41.35	60.90	61.54	70.00
Aligned reader, lossless integer-to-text conversion								
Official task-specific	60.80	82.86	92.86	43.33	33.08	58.65	75.64	63.33
Mem0 unified	66.60	91.43	92.86	60.00	43.61	60.90	76.92	70.00
Evidence selection and answer review								
Official task-specific	66.00	82.86	98.21	36.67	44.36	66.17	75.64	60.00
Mem0 unified	71.40	92.86	98.21	53.33	51.13	68.42	79.49	80.00

Judge models. We score the same Qwen3-4B-Instruct-2507 answers with five Judge services, holding the public Mem0 prompt, temperature, output limit, and disabled-thinking request setting fixed. Table 13 reports overall accuracies from 66.00% to 71.40%, a spread of 5.40 percentage points. Relative to DeepSeek-V4-Flash, Qwen3.6-35B-A3B, GPT-6-sol, GPT-6-luna, and Claude-fable-5 change 20, 25, 27, and 28 correct judgments to incorrect, respectively; the reverse changes occur on 2, 1, 1, and 1 questions. Model names are service identifiers; underlying model identities are not independently verified. These results measure sensitivity to Judge services on fixed answers.

Table 13: LongMemEval accuracy (%) for fixed answers under five Judge services using the same public judging prompt. Abs.* denotes the overlapping abstention subset.

Judge	Overall	SS-U	SS-A	SS-P	Multi	Temp	Update	Abs.*
DeepSeek-V4-Flash	71.40	92.86	98.21	53.33	51.13	68.42	79.49	80.00
Qwen3.6-35B-A3B	67.80	84.29	98.21	40.00	46.62	66.92	79.49	76.67
GPT-6-sol	66.60	84.29	96.43	33.33	45.86	66.17	78.21	76.67
GPT-6-luna	66.20	82.86	96.43	33.33	45.86	65.41	78.21	76.67
Claude-fable-5	66.00	81.43	96.43	33.33	46.62	64.66	78.21	73.33

Official task-specific templates. The English blocks below form the actual prompts. Ordinary QA and multi-session questions use C+F; temporal questions use C+T+F; knowledge updates use U+F; preferences use P. Abstention takes precedence and uses A. The `{}` fields receive the question, reference or rubric, and model response in order.

C: ordinary QA and multi-session questions

```
I will give you a question, a correct answer, and a response from a
model.  Please answer yes if the response contains the correct
answer.  Otherwise, answer no.  If the response is equivalent to the
correct answer or contains all the intermediate steps to get the
correct answer, you should also answer yes.  If the response only
contains a subset of the information required by the answer, answer
no.
```

T: temporal addition

```
In addition, do not penalize off-by-one errors for the number of
days.  If the question asks for the number of days/weeks/months,
etc., and the model makes off-by-one errors (e.g., predicting 19 days
when the answer is 18), the model's response is still correct.
```

U: knowledge updates

```
I will give you a question, a correct answer, and a response from a
model.  Please answer yes if the response contains the correct
answer.  Otherwise, answer no.  If the response contains some
previous information along with an updated answer, the response
should be considered as correct as long as the updated answer is the
required answer.
```

F: shared input and output format

```
Question:  {}
Correct Answer:  {}
Model Response:  {}
Is the model response correct?  Answer yes or no only.
```

P: preferences

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

A: abstention

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

Mem0 unified template. All questions use the following complete English prompt, taken from `JUDGE_PROMPT` at Mem0 commit `7ba1bd3`. Named fields receive the question, reference or rubric, and model response.

Mem0 Unified Judge Prompt

```
I will give you a question, a correct answer (or rubric), and a model
response.  Decide whether the model response is correct.  Answer yes
or no only.
CORE PRINCIPLE -- Semantic equivalence:  Judge by MEANING, not exact
words.  Answer "yes" if every concept in the correct answer is
```

```
addressed in the response, even with different vocabulary, more
specific terms, or restructured phrasing.
Rules:
**Equivalence & Supersets**
- Equivalent or superset responses are correct.  Extra details are
fine unless proven to be factually wrong.  Extra qualifiers are fine
unless proven to be wrong.  E.g., "a blue dress and a matching
necklace" is correct when the answer is "a blue dress."
- If a response captures the most specific part (exact
item/place/name) but omits a broader container, it's correct.
- Same factual meaning with different phrasing = correct (e.g., "No,
you did not visit with a friend" ~ "You didn't mention going with
anyone").
**Lists & Compound Terms**
- For list answers, match each item by semantic meaning.  A concept
is covered if restated via synonyms, sub-concepts, or related terms.
Adding methodological detail or rewording verbs to near-synonyms is
acceptable.
- A broad term like "A and B significance" is covered if the response
addresses the topic area through related specific terms, even without
naming each component literally.
- If some items as listed as "or"s, "maybe"s and potential answers,
it's okay if the answer does not include those.
- If two items in a list achieve the same purpose, listing just one
of them is fine.
IMPORTANT: The "anti-preference" items are very specific!
Eg.  Someone "not interested in general AI topics" could be very
interested in specific AI topics in general AI *conferences*; those
are not the same thing and should be accepted!  topics != conferences
**Numbers & Precision**
- Hedging ("at least 3", "approximately") is fine if the core number
matches.  A range that includes the correct answer is correct.
Generally, if the user themself would be satisfied by the response,
it is acceptable.
- More precise answers are correct:  "22 days" matches "3 weeks";
"over $270" matches "$270."; "9 1/2 months" matches "9 months";
- Rough answers are correct:  "about nine months" ~ "9 months; "8
months and 20 days" matches "9 months";
- Off-by-one errors on days/weeks/months are acceptable.
- Approximate unit conversions are equivalent:  "14 weeks" ~ "3
months", "6 months" ~ "half a year."
- Round time ranges generously:  7 months and 16 days ~ 8 months.
- Notes instead of chords are acceptable when justified
**Dates & Temporal**
- Date format variations are equivalent:  "February 1st" = "Feb 1,
2023" = "on February 1."
- Same-day event ordering swaps are acceptable.
- Outdated info alongside the correct updated answer is acceptable if
the current value is identified.
- "recent" is upto 6 years ago, which means 2017+
- References like "last weekend", "last Wednesday", etc.  are
imprecise - people sometimes mean the weekend/Wednesday before the
latest one if they're near it.  "Last 3 months" can include boundary
days of the 4th month back.  "Last month" includes the current month
so far.  Be flexible with such timestamps
**Counting Edge Cases**
- If correct answer is "0" or "nothing found," model saying "not
enough information" is also correct.
- Similarly, If correct answer is "not enough information", model
saying "0" or "nothing found," is also correct.
**Preference/Personalization Rubrics** (apply in order):
1.  Correct if the response demonstrates awareness of user's personal
context (preferences, habits, interests).  Need not satisfy every
rubric point.
```

```
2. Primary criterion: do main suggestions align with what the user
WANTS?
3. Anti-preferences: evaluate the OVERALL thrust, not keyword
scanning. If the response largely suggests correct options, minor
incidental references to "not-preferred" things are fine.
4. Mentioning a phone app as a MEANS to a preferred activity (e.g.,
meditation app for sleep) is not "suggesting phone use." Judge by the
activity, not delivery mechanism.
5. "May not prefer" = mild preference, not hard prohibition.
Secondary/context-dependent inclusion is fine.
6. Explicit acknowledgment of anti-preferences (e.g., "keep screens
off") strengthens correctness.
7. Context-dependent suggestions are acceptable (reading is fine on
a bus even if rubric flags visual attention activities). Adjacent
genres alongside preferred ones are additive, not contradictory.
**Abstention Matching**
- If correct answer = unanswerable, any phrasing conveying "not
enough information" is correct. Providing partial context while
concluding the question can't be answered = correct abstention.
FINAL CHECK: Before answering "no," ask: is the concept truly
absent, or expressed via equivalent terminology? Only answer "no" if
a core concept is entirely unaddressed.
Question: {question}
Correct Answer: {answer}
Model Response: {response}
Is the model response correct? Answer yes or no only.
```

A.7 EVIDENCE ORGANIZATION AND ANSWER REFINEMENT

Once relevant memories have been retrieved, the reader must locate facts across conversations, interpret temporal relationships, and combine evidence into an answer. We examine how evidence presentation, selection, and answer review affect this process on a fixed 120-question LongMemEval development set. All configurations share the base retrieved evidence pool, Qwen3-4B-Instruct-2507 reader, and official task-specific DeepSeek-V4-Flash judging protocol described in Appendix A.6.

Table 14: Evidence organization and answer refinement on the development set. Accuracy is a percentage; all configurations retain the 120-question denominator, including failed answers. Bold and underline mark the highest and second-highest scores.

Configuration	Time	Select	Checklist	Window	Dedup	Review	Correct	Acc.
All retrieved evidence	–	–	–	–	–	–	75/120	62.50
Time index	✓	–	–	–	–	–	74/120	61.67
Selection + retention	✓	✓	–	–	–	–	<u>80/120</u>	<u>66.67</u>
Question checklist	✓	✓	✓	–	–	–	77/120	64.17
Message windows	✓	✓	–	✓	–	–	78/120	65.00
Identity deduplication	✓	✓	–	–	✓	–	78/120	65.00
Post-selection review	✓	✓	–	–	–	✓	**84/120**	**70.00**

Table 14 lists the active mechanisms. Time adds source dates and their offsets from the question date. Select combines grouped evidence selection with retention of the original top-4 retrieved blocks. Checklist adds question requirements; Window selects message-level context; Dedup replaces repeated native message identities with references; Review checks a draft against the selected evidence. Context length, model-call count, and output budget vary across configurations.

Adding a time index changes accuracy from 62.50% to 61.67%. Source dates help locate statements in the conversation timeline, but the reader must still distinguish when a statement was made from when its described event occurred. Explicit source-time structure alone provides no aggregate gain on this development set.

Selection with top-4 retention reaches 66.67%. It narrows the evidence to be read while retaining high-ranked blocks from the original retrieval, limiting reliance on a single selection decision. Adding a question checklist, message windows, or identity-based deduplication yields 64.17%, 65.00%, and 65.00%, respectively. These configurations emphasize question requirements, local context, and redundancy control, yet none improves on the simpler selection-and-retention combination here.

Post-selection review gives the highest observed accuracy, 70.00%. The reader first drafts an answer from the selected evidence and then performs one review for every question, without using Judge feedback or known correctness to decide when to review. Relative to selection with retention, paired scoring records show 5 wrong-to-correct and 1 correct-to-wrong transitions. The result favors further study of answer refinement after evidence selection, while showing that review can also introduce regressions.

A.8 STATE AND EVIDENCE COMPONENT ANALYSIS

Table 4 reports TimlMemBench raw-conversation behavior scores with GPT-5.4-mini, removing versioned state, Current View, final eligibility rechecking, or all three. Chinese and English results are reported separately using their axis-specific scoring.

Removing versioned state produces the largest Supersede decrease, 8.6 points in both languages, and lowers Revoke by 5.8 and 6.0 points. Removing Current View lowers Supersede by 4.5 and 4.3 points. Final eligibility rechecking has smaller individual effects, mainly on scope, provenance, and use control. Removing all three lowers Supersede by 12.6 and 12.9 points and reduces every axis.

Table 15: Cumulative LoCoMo configurations with GPT-5.6-terra. T: temporal evidence processing; C: coverage planning; S: write-time state storage; V: Current View consumption; E: dedicated answer-evidence selection; M: explicit multi-hop decomposition; U: unified message-source contract.

Configuration	T	C	S	V	E	M	U
A	–	–	–	–	–	–	–
B	–	✓	✓	✓	–	–	–
C	✓	✓	✓	✓	–	–	–
D	✓	✓	✓	✓	✓	✓	–
Full	✓	✓	✓	✓	✓	✓	✓

Table 16: LoCoMo F1/B1 for cumulative configurations (%). Bold and underline mark the highest and second-highest distinct scores per column.

	Single-hop		Multi-hop		Open-domain		Temporal	
Configuration	F1	B1	F1	B1	F1	B1	F1	B1
A	63.14	57.29	29.09	23.12	25.87	21.37	16.13	12.26
B	64.63	58.83	30.50	24.55	<u>29.95</u>	<u>25.19</u>	22.08	17.68
C	59.31	54.09	24.45	19.63	22.91	18.58	57.82	<u>54.05</u>
D	<u>65.87</u>	<u>59.48</u>	<u>45.24</u>	<u>38.51</u>	25.49	21.50	**61.32**	**54.77**
Full	**69.11**	**62.13**	**52.30**	**45.57**	**30.03**	**25.48**	<u>58.32</u>	53.01

A–B adds coverage planning, write-time state storage, and Current View together, improving all four F1 categories. B–C includes temporal evidence processing; temporal F1 rises by 35.74 points while the other three categories decrease. C–D adds dedicated evidence selection and multi-hop decomposition together, raising multi-hop F1 by 20.79 points. D–Full adds the unified message-source contract: single-hop, multi-hop, and open-domain F1 rise by 3.24, 7.06, and 4.54 points, while temporal F1 falls by 3.00 points.

A.9 GOVERNANCE CHECKS AND PROCESSING EFFICIENCY

Evidence eligibility checks. We evaluate final post-selection eligibility checks in controlled source-invalidation scenarios. Paired conditions share the same sources and questions, with GPT-6-luna as the answer model. Enabling final source checking reduces violating answers from 12 to 0 across 12 source-invalidation conditions, while preserving all four authorized controls.

We also examine the interaction between upstream and final scope checks. In four unauthorized conditions with upstream checking disabled, final scope checking blocks the state from entering the model request. Removing only the final scope check exposes out-of-scope evidence in all four conditions. Evidence summaries remain identical across eight authorized controls.

Table 17: Evidence eligibility check results.

Measure	Check enabled	Check removed
Invalid-source answers ↓	**0/12**	12/12
Authorized controls retained ↑	**4/4**	4/4
Out-of-scope requests after upstream omission ↓	**0/4**	4/4

Scope and temporal constraints. In a native state-update scenario, we cross query dates, authorized scopes, and check settings. Scope checks reject unauthorized reads; interval filtering selects the date-appropriate state for authorized queries. Queries before the update return the old value, and queries after it return the new value. Final scope checking also excludes unauthorized states admitted upstream, enforcing eligibility from reading through request construction.

Scope-filtering efficiency. We pool the 18 original worlds from an independent nine-axis sample and define access scope by world membership. We measure Reader token usage and local processing time with questions, model settings, and evidence budgets fixed. Scope filtering reduces out-of-scope evidence entering context from 335 items to 0 and reduces the content retrieved, assembled, and passed to the Reader.

Table 18: Reader token usage and local processing time with scope filtering.

Measure	Enabled	Disabled	Reduction
Reader input tokens	**113,672**	230,089	50.6%
Reader input and output tokens	**119,170**	237,222	49.8%
Median local processing time	**18.95 ms**	160.97 ms	88.2%

We time retrieval, evidence budgeting, eligibility checks, and request assembly. Each question receives two warmups and 20 alternating paired measurements, yielding 360 pairs. Scope filtering reduces the median time for all 18 questions, with assembled requests matching the corresponding frozen requests.

Current View read efficiency. For 16 applicable current-state queries, Current View and historical reads at the same timestamp return identical record sets. Median local read times are 2.31 ms and 3.50 ms, respectively, demonstrating the processing efficiency of the dedicated current-state read path.

A.10 WORKED EXAMPLE: STATE EVOLUTION AND GOVERNED EVIDENCE USE

This example illustrates how source evidence, versioned state, and Current View support changing conversational preferences.

A changing preference. All statements concern the same user's drink preference within the same authorized scope.

1. **June 1—persistent preference.** `I usually go for sugar-free drinks.` A persistent preference is recorded with a reference to this statement.
2. **June 10—temporary exception.** `I'm training a lot this week. Low-sugar drinks are fine through June 16.` A time-bounded exception overlays the persistent preference through the specified date.
3. **June 17—expiry.** With no further update, the temporary exception expires and the sugar-free preference becomes visible again in Current View.
4. **June 20—preference update.** `I prefer low-sugar drinks now. Let's go with those from now on.` The explicit change supports a new persistent version and closes the earlier version's validity interval, retaining both sources.
5. **June 25—preference relaxation.** `I'm not particular about sugar anymore--sugar-free or low-sugar is fine.` The current preference now admits either option. The earlier low-sugar preference no longer supplies an exclusive recommendation criterion.

State evolution and answer evidence. At each of the five dates above, we ask the same question: `Recommend a drink matching my preferences.` Relation extraction, read planning, and answering use GPT-6-luna. We denote the state versions derived from the four user statements by v_1–v_4: v_1 is the initial sugar-free preference, v_2 the temporary exception permitting low-sugar drinks, v_3 the updated persistent low-sugar preference, and v_4 the relaxed preference accepting either option. Each version retains its source reference. When the temporary exception expires, the current state returns to v_1 without creating a new version. Table 19 shows the current state and the version cited by each answer. Inputs and answers in this execution are in Chinese; English renderings are translations.

Table 19: Drink-preference state evolution and answer citations. Versions v_1–v_4 correspond to the four source statements.

Date	Preference change	Current state	Cited version
June 1	Sugar-free preference established	v_1: usually chooses sugar-free drinks	v_1
June 10	Low-sugar temporarily accepted	v_2: permits low-sugar drinks temporarily; persistent preference retained	v_2
June 17	Exception expires; no new statement	v_1: sugar-free preference restored	v_1
June 20	Preference updated to low-sugar	v_3: new persistent low-sugar preference	v_3
June 25	Sugar preference relaxed	v_4: either sugar-free or low-sugar	v_4

The temporary exception preserves the original persistent preference, allowing it to reappear on expiry without another user statement. The two subsequent updates each create a persistent version and close the previous version's validity interval. All five answers cite the state evidence for the corresponding date. On June 17, the answer recommends sugar-free sparkling water and cites the restored v_1. On June 25, it recommends sugar-free oolong tea, explaining that it `matches your current preference allowing either sugar-free or low-sugar drinks`, and cites v_4. The trace connects changing user statements to state versions and traceable answer evidence.

Current and historical questions. For `What drink would suit my preferences?`, the relevant state permits low-sugar drinks on June 12, returns to sugar-free on June 18, and favors low-sugar after the June 20 update. After June 25, either option is consistent with the expressed preference.

A historical question asked on June 22—`What was my preference on June 18?`—uses the version valid at the requested time and recovers the sugar-free preference. Source references provide the evidence for each transition.

From stored evidence to answer context. After June 25, retrieval may still find the earlier statement favoring low-sugar drinks. Its source remains part of the history, while its temporal status determines how it can support the answer. A current recommendation uses the relaxed preference; a historical question can use the earlier version when its time and access conditions match. Versioned state records these changes, Current View presents the current state, and final eligibility checks govern the evidence admitted for the request.

The last statement updates the preference; it requests neither deletion of the conversation nor withdrawal of permission to use its history. If an utterance leaves the update target or time unclear, the system retains its source for user review before applying the corresponding state change.

A.11 EVIDENCE USE AFTER POST-SELECTION REVOCATION

We examine whether a revocation after candidate selection takes effect before answer-context assembly. Each condition uses an independent SQLite database, the normal revocation interface, and a mock Reader that captures the complete request.

Input and procedure. The source message and its memory record contain the following Chinese statement, translated here: *I usually go to the book club at Qinghe Bookstore on Wednesday evenings.* The question is: *When do I usually go to the book club?* The initial memory is active and belongs to the request's authorized scope. We read the record into a fixed candidate pool and retain its cached object.

At a synchronization point between evidence selection and context assembly, we either leave the record unchanged or revoke it through the normal interface. Revocation changes the persisted status to `deleted`, while the cached object remains `active`. We then resume the answer pipeline with the final persisted-source recheck enabled or disabled. The enabled configuration rereads the persisted source status; the disabled configuration uses the cached status. Other in-memory filters remain enabled. Each condition produces one captured Reader request.

Results. We inspect every message in the request for the full original statement and the Chinese phrases corresponding to *Wednesday evenings* and *Qinghe Bookstore*. All three checks agree (Table 20).

Table 20: Post-selection revocation and evidence inclusion in the Reader request.

Operation	Final recheck	Cached status	Persisted status	Included
Unchanged	Enabled	active	active	Yes
Unchanged	Disabled	active	active	Yes
Revoked	Enabled	active	deleted	No
Revoked	Disabled	active	deleted	Yes

Without revocation, both configurations retain the original statement with valid evidence ID `E1`. After revocation, the configuration without rechecking still includes it as `E1`; the enabled configuration removes it and produces the following evidence block:

Evidence Block After Revocation

```
[Prompt-safe Memory Evidence]
No prompt-safe memory evidence was selected.
[Valid Evidence IDs]
[]
```

In this case, the final source recheck detects revocation after candidate selection and prevents stale cached evidence from entering the answer context, while preserving evidence whose eligibility remains unchanged.