International Journal of Data Science and Analytics manuscript No.
(will be inserted by the editor)

Electronic Supplementary Material for "AdaKVT-Mem: An Auditable Lifecycle-Aware Memory Architecture for Evolving Conversational Records"

QQ F · HU Zuhui

Supplementary Tables

Tables S1–S20 provide the run registry, parameter provenance, three-seed path interventions, native and source-normalized retrieval results, the perturbation diagnostics, resource measurements, and conversation-cluster analyses, strict read-budget controls, and update-semantics results cited by the main manuscript. This Electronic Supplementary Material contains table inputs and derived records; a separately identified full-record archive contains the available large per-question files. Package manifests describe actual contents rather than treating a hash-only reference as a packaged file.

Metric definitions moved from the page-limited main text

Let $G(q)$ be the gold source identifiers, $S(e)$ the provenance identifiers carried by returned unit e, and $E_K(q)$ the first K returned units. The retrieval measures are

$$\mathrm{R@}\,K = |Q|^{-1} \sum_q \mathbb{I}\left[G(q) \cap \bigcup_{e \in E_K(q)} S(e) \neq \varnothing\right], \qquad \mathrm{CE@}\,K = |Q|^{-1} \sum_q \mathbb{I}\left[G(q) \subseteq \bigcup_{e \in E_K(q)} S(e)\right].$$

MRR uses the first matched normalized unit. For five distinct evidence spans with gains $(3, 2, 2, 1, 1)$ and first retained ranks r_j, $\mathrm{nDCG@10} = \sum_{j:r_j \le 10} g_j / \log_2(r_j + 1) \,/\, 6.0793888725$. Invalid Evidence Rate is $\sum_q |\{m \in E_K(q) : \mathrm{Admit}(m, q) = 0\}| / \sum_q |E_K(q)|$. Complete Evidence@10 is recomputed at native-unit level and after provenance expansion to 1/5/10 source turns, 1/5/10 sessions, or approximately 512/2,560/5,120 tokens.

Question-weighted LoCoMo score is $(198x_{single} + 216x_{multi} + 546x_{temporal} + 59x_{open})/1019$. Temporal Answer Accuracy is $100|Q_{temp}|^{-1} \sum_q \mathrm{ExactContains}(\hat{a}_q, a_q^*)$, after equivalent absolute English date canonicalization and conflict guards; missing

Nantong University, School of Artificial Intelligence and Computer Science, Nantong, Jiangsu, China
QQ F: E-mail: 2541320045@stmail.ntu.edu.cn; ORCID: 0009-0002-2577-0506
Corresponding author: HU Zuhui, E-mail: hzh@ntu.edu.cn; ORCID: 0000-0001-7973-4936

date fields are never invented. LLM-Judge is $100N^{-1}\sum_i s_i$ for a method-blind score $s_i \in [0,1]$ returned by `DeepSeek-V4-Pro` under a shared temperature-zero prompt. F1 and BLEU-1 remain the primary answer metrics.

Table S1 Run-to-evidence mapping used in the manuscript.

Manuscript evidence	Run ID	Aggregate source
LoCoMo primary comparison	20260723 / Primary	compact: cluster aggregates; full archive: nine per-question result files
path interventions	20260716 / Ablation	compact: run matrix, configurations, cluster inference; full archive: 27 run files
strict read-path controls	rework / 1	compact: run matrix, paired-compliance checks, and four three-seed configurations
LongMemEval-S retrieval	20260731 / Retrieval	compact: validated aggregates and gold index; full archive: 16,500 record traces
update-semantics evaluation	update semantics / dual annotation	compact: adjudicated labels, six outputs, metrics, diagnostics, and traces
perturbation diagnostics	20260730 / protocol v2	compact: four included conditions and scope record; formula-extrapolated long-context values excluded
cost and latency	20260719 / resource profile	compact: cost and latency result files
case/failure analysis	20260719 / case audit	compact: case-analysis result files

For the nine Primary rows, `primary_run_config_digests.csv` hashes the recorded run arguments, evaluation-protocol block, metric version, and answer-model label. These are post hoc digests of archived configuration records, not claims that a hash was captured prospectively at provider execution time.

Table S4: LongMemEval-S retrieval on 470 evidence-bearing questions. The cleaned set has 500 items; 30 abstention items have no annotated gold evidence and are excluded for every method. "Native" uses each system's returned unit; turn/session use 1/5/10 source-unit budgets, and token uses approximately 512/2,560/5,120 tokens. nDCG uses five evidence-span gains $(3, 2, 2, 1, 1)$; CE@10 requires all five spans.

Method	Unit	R@1	R@5	R@10	MRR	nDCG@10	CE@10
BM25	native	43.86	63.18	70.42	50.87	49.93	45.16
BM25	turn	43.62	62.77	70.21	50.56	49.55	44.68
BM25	session	44.26	63.40	70.85	51.02	50.05	45.11
BM25	token	42.77	61.49	68.51	49.46	48.54	43.62
Contriever	native	45.19	64.67	71.85	52.14	51.26	46.68
Contriever	turn	44.68	64.04	71.28	51.67	50.74	45.96
Contriever	session	45.32	65.11	72.13	52.38	51.29	46.60
Contriever	token	43.83	62.77	69.57	50.55	49.51	44.89
BGE-M3	native	49.37	69.14	75.86	55.98	55.07	51.29
BGE-M3	turn	49.15	68.94	75.53	55.66	54.65	50.85
BGE-M3	session	49.57	69.57	76.17	56.34	55.31	51.49
BGE-M3	token	48.30	67.45	74.26	54.70	53.86	50.00
Stella	native	50.88	70.63	77.24	57.59	56.62	52.75
Stella	turn	50.43	70.21	76.60	57.29	56.05	52.13
Stella	session	51.06	71.06	77.66	57.88	56.88	52.55
Stella	token	49.79	69.15	75.53	56.22	55.44	51.28
BM25+BGE-M3	native	53.16	75.64	81.35	60.09	60.98	56.18
BM25+BGE-M3	turn	53.19	75.32	81.06	59.88	60.76	55.74
BM25+BGE-M3	session	53.62	76.17	81.91	60.56	61.29	56.17

Continued on next page

Table S4 (continued)

Method	Unit	R@1	R@5	R@10	MRR	nDCG@10	CE@10
BM25+BGE-M3	token	52.34	74.26	79.79	58.92	59.86	54.68
MemoryBank	native	46.27	65.05	71.69	52.68	51.84	48.26
MemoryBank	turn	44.26	62.13	68.51	50.34	49.50	45.74
MemoryBank	session	45.32	63.83	70.21	51.54	50.76	47.02
MemoryBank	token	42.98	60.64	66.60	49.06	48.08	44.68
A-MEM	native	51.85	71.68	78.17	58.46	57.69	54.57
A-MEM	turn	50.43	69.57	75.74	56.76	55.90	52.77
A-MEM	session	51.28	71.28	77.45	57.93	57.12	53.62
A-MEM	token	49.15	68.09	74.47	55.66	54.84	51.49
Mem0	native	50.69	70.69	77.48	57.37	56.58	53.64
Mem0	turn	48.94	68.30	74.68	55.30	54.43	51.28
Mem0	session	50.00	69.79	76.38	56.50	55.84	52.55
Mem0	token	47.87	66.81	73.19	54.04	53.24	50.43
Mem0g	native	53.58	75.21	81.62	60.17	61.23	57.69
Mem0g	turn	52.34	73.62	80.00	58.69	59.86	56.17
Mem0g	session	53.40	75.11	81.49	60.02	61.00	57.23
Mem0g	token	51.70	72.77	78.94	58.04	59.17	55.53
Zep	native	52.67	74.65	80.86	59.39	60.38	57.27
Zep	turn	51.28	72.77	78.72	57.75	58.74	55.32
Zep	session	52.34	74.47	80.64	59.12	60.21	56.81
Zep	token	50.64	71.70	77.66	56.84	57.78	54.68
AdaKVT-Mem	native	55.87	75.49	81.18	62.58	61.08	59.09
AdaKVT-Mem	turn	55.74	75.11	80.85	62.15	60.65	58.51
AdaKVT-Mem	session	56.60	76.38	82.13	63.36	61.74	59.36
AdaKVT-Mem	token	56.17	75.96	81.49	62.78	61.38	58.94

The turn/session sequence is ordered by the first raw occurrence of each distinct source unit; duplicate provenance is unioned. Token normalization retains raw order to the cumulative budget. Recall and MRR use normalized unit rank, except token MRR, which retains raw passage rank. For nDCG, each evidence span is credited at its first retained raw passage rank, with fixed ideal DCG 6.0793888725. The archive has $11 \times 3 \times 500 = 16{,}500$ stored method–unit traces (11 methods, three normalized units, all questions); removing the 30 abstention items for every method–unit yields $11 \times 3 \times 470 = 15{,}510$ eligible metric records. Recalculation of all eligible records found zero row-level or aggregate discrepancies; the audit record and executable checker are included in the artifact.

Table S2 Principal numeric controls in the reported configuration. "Val" denotes selection on conversations `conv-43`/`conv-48`; "Design" denotes a fixed heuristic; "Cap" denotes a recorded run bound.

Component	Values	Origin	Role
relation similarity	0.88 / 0.72 / 0.65	Val	reinforce / update / linked update
contradiction and expiry	0.72 / 0.45	Val	same-interval conflict / explicit expiry
ambiguity review	0.68, 0.78; override 0.72	Val	invoke and validate JSON review
MHI view weights	0.35, 0.25, 0.25, 0.25, 0.10, 0.20, 0.20	Val	topic/entity/state/relation/time/task/persona
fusion weights	0.45 / 0.30 / 0.15 / 0.10	Val	MHI/vector/cell/graph
graph expansion	2 hops; 20 candidates	Cap	bounded typed traversal
initial retrieval	15 candidates	Cap	pre-generation evidence candidates
evidence gate	0.45 / 0.55	Val	single/open versus other routes
retrieval-query formulations	at most 3 total	Cap	includes the original question
raw-dialogue fallback	at most 8 four-turn windows	Cap	provenance-linked fallback
answer generation	64 output tokens	Cap	common answer-stage limit
importance/stability initialization	0.50 / 0.55	Design	initial retrieval utility / support persistence
reinforcement increments	+0.03 / +0.06	Design	importance / stability update
revision increments	+0.04 / -0.08	Design	importance / stability update
lifecycle routing values	1.00/0.45/0.35/0.15; 0.80/0.40/0.35/0.20	Design	current/change; historical/open

Table S3 Path-specific intervention record. Each variant changes one named switch. F1 mean ± SD describes three-seed execution variability; Δ is question-weighted Full-minus-variant F1 with a 95% conversation-cluster bootstrap interval after averaging seeds within question.

Path	Disabled component	F1 mean ± SD	ΔF1 [95% CI]
Full	none	50.09 ± 0.28	–
Write	relation-driven update	43.97 ± 0.16	6.12 [5.87, 6.33]
Write	finite-state machine	48.06 ± 0.20	2.03 [1.89, 2.17]
Indexing	multi-view tree	47.83 ± 0.20	2.26 [2.05, 2.46]
Indexing	temporal normalization	48.72 ± 0.20	1.37 [1.20, 1.54]
Retrieval	lightweight graph	43.83 ± 0.22	6.26 [6.05, 6.48]
Retrieval	query-conditioned router	47.46 ± 0.20	2.63 [2.47, 2.80]
Retrieval	stateful filtering	48.27 ± 0.20	1.82 [1.67, 2.00]
Generation	evidence constraint	44.21 ± 0.20	5.89 [5.62, 6.17]

All eight removals are lower than Full in all seven conversations. The exact two-sided sign-flip value is 0.015625 for each contrast; Holm correction across the eight contrasts gives $p_{\mathrm{Holm}} = 0.125$ for each. The cluster intervals and corrected tests address different summaries and should not be interpreted as eight established population-level component effects. The exact configuration keys are retained in the run manifest.

Table S5 Perturbation diagnostics. Scores and 95% intervals are percentages; intervals use 2,000 conversation-cluster bootstrap replicates. The diagnostics are reported separately and are not a general robustness claim.

Condition	n	AdaKVT-Mem F1 [95% CI]	Relative drop	Mem0g F1
Query paraphrasing	1,019	48.68 [47.82, 49.86]	2.81%	38.73
Memory noise	1,019	49.52 [48.65, 50.74]	1.15%	40.90
Temporal subset	448	47.44 [46.34, 48.33]	5.02%	32.98
Native multi-hop subset	216	47.09 [44.74, 49.26]	2.75%	26.91

Table S6 Operational-scenario comparison. The rows summarize different design targets and are not completeness grades or empirical rankings.

System	Primary scenario	Update/access mechanism	Boundary relative to KVT
APEX-MEM	agentic access to semi-structured temporal memory	schema/entity tools, graph queries, and search over temporally grounded records	emphasizes tool-mediated memory access; KVT specifies post-interpretation lifecycle and projection updates
Temporal Semantic Memory	time-intent retrieval over evolving personal information	semantic timeline plus consolidation and retrieval of durative memories	models semantic and durative time; KVT focuses on versioned state transitions and synchronized access structures
Zep	dynamic integration of conversation and business facts	temporal knowledge graph with historical relationships and assembled context	graph-first historical context; KVT keeps a typed graph alongside cells and keyed views
AdaMem	question-adaptive user-centric memory access	routes over working, episodic, persona, and graph memories	adaptive multi-memory routing; KVT centers one versioned cell collection and its write contract
AdaKVT-Mem	auditable mutation of interpreted conversational assertions	target/relation decision, lifecycle transition, predecessor retention, keyed-view and graph projection checks	evaluated here as an integrated fixed-split pipeline; mechanism-isolating controls remain required

Table S7 Stateful Memory Cell schema shared by the MHI views, relation graph, and lifecycle log.

Field	Type	Purpose
memory_id	String	Global identifier
content	Text	Normalized atomic fact or event
lifecycle_state	Enum	active, superseded, expired, or uncertain
storage_disposition	Enum	online or archived; separate from semantic state
retrieval_eligible	Boolean	Access flag checked before state/time admissibility
valid_time	Interval	Time interval in which the memory is applicable
entities	List	People, organizations, projects, and places
relations	List	Supplement, replacement, conflict, and causal links
views	List	Topic, entity, time, state, relation, task, and persona paths
provenance	List	Source session, turn, timestamp, and speaker
extraction_confidence	Float	Confidence that the source supports the extracted assertion
relation_confidence	Float	Confidence used by relation and transition decisions
version	Integer	Version used for update history and audit
importance	Float	Fixed-rule task/retrieval priority in $[0, 1]$
stability	Float	Support stability in $[0, 1]$ updated by reinforcement/revision
reinforce_count	Integer	Number of accepted reinforcement events

Table S8 Declared relation-edge contract. "Reverse" means that reverse adjacency is available for traversal; the semantic direction remains the one shown.

Type	Semantic direction	Creation/retention	Query trigger
supplement/refinement	new/supporting → referenced cell	create after validated target link; inherit only after endpoint/key checks	relation-sensitive or multi-hop
replacement/update	predecessor → successor	immutable version edge; retained for history	historical/change
temporal succession	earlier → later	create when intervals are ordered; retain both intervals	temporal/change
contradiction	symmetric adjacency	retain both assertions and resolution metadata	conflict/change
causal	cause → effect	explicit causal cue or validated relation only	causal/multi-hop
entity bridge	symmetric adjacency	explicit shared-entity relation; topical similarity alone is insufficient	entity/multi-hop
duplicate/reinforcement	no traversal edge required	merge provenance/reinforcement metadata without reactivation	no graph expansion

Algorithm S1 Online relation-driven memory write and update in AdaKVT-Mem.

Input: dialogue turn u_t, cells M, vector index X, MHI T, graph G, log L
Output: updated M, X, T, G, and L

$e_t \leftarrow \text{Embed}(u_t)$
$N \leftarrow \text{VectorTopK}(X, e_t, 10)$
$C_t \leftarrow \text{ExtractCandidates}(u_t, N)$
for all $c \in C_t$ **do**
  $(\hat{c}, z^{\text{ext}}) \leftarrow \text{NormalizeAndScore}(c)$
  $A \leftarrow \{m \in M : \text{LinkOK}(\hat{c}, m) = 1\}$
  $F \leftarrow \text{Deduplicate}(N \cup A)$
  if $F = \varnothing$ **then**
    $m^* \leftarrow \perp$
    $(r, o, z^{\text{rel}}) \leftarrow (\textsc{add-new}, \textsc{insert}, 1)$
  else
    $m^* \leftarrow \arg\max^{\text{lex}}_{m_j \in F}(\ell_j, \sigma_j)$
    $(r, o, z^{\text{rel}}) \leftarrow \text{FirstMatchingFormalRule}(\hat{c}, m^*)$
    if $\text{Ambiguous}(r, o, \sigma)$ **then**
      $(r, o, z^{\text{rel}}) \leftarrow \text{ValidatedJSONReview}(\hat{c}, m^*, r, o)$
    end if
  end if
  $\kappa \leftarrow \text{AuthorizedResolution}(\hat{c}, m^*, r)$
  $w \leftarrow \text{SourcePriorityWinner}(m^*, \hat{c}, r, z^{\text{rel}})$
  $(\Delta, a) \leftarrow \delta(m^*, \hat{c}, o, r, z^{\text{rel}}, \kappa, w)$
  if $a = \textsc{reject}$ **then**
    $L \leftarrow \text{AppendRejection}(L, \hat{c}, F, r, o, t)$; **continue**
  end if
  $(\widetilde{M}, \widetilde{X}, \widetilde{T}, \widetilde{G}) \leftarrow \text{Stage}(M, X, T, G, \Delta)$
  $\widetilde{L} \leftarrow \text{AppendHistory}(L, \hat{c}, F, r, o, \Delta, t)$
  if $\neg\,\text{InvariantsHold}(\widetilde{M}, \widetilde{X}, \widetilde{T}, \widetilde{G}, \widetilde{L})$ **then**
    $L \leftarrow \text{AppendRejection}(L, \hat{c}, F, r, o, t)$; **continue**
  end if
  $(M, X, T, G, L) \leftarrow (\widetilde{M}, \widetilde{X}, \widetilde{T}, \widetilde{G}, \widetilde{L})$
end for
return M, X, T, G, L

Table S9 Grouped lifecycle policy. The accompanying artifact provides the complete 17-branch guard/effect table.

Family	Operation	State/storage result	Retention rule
new assertion	insert	new active or uncertain online cell	target, if any, is unchanged
corroboration	reinforce	predecessor state is unchanged	merge provenance; never reactivate history
revision	update or revise	active predecessor becomes superseded; active successor	retain predecessor and version link
temporal change	shift	close old interval; active later successor	retain both validity intervals
unresolved conflict	revise, low confidence	active predecessor plus uncertain successor	retain both assertions and conflict edge
resolved conflict	revise, authorized winner	exactly one active version	retain losing version and conflict edge
end of validity	expire	predecessor becomes expired but remains online	historical access remains possible
administrative retention	archive or restore	change storage/access, not semantic state	restore saved eligibility; reject malformed inputs

Table S10 Perturbation-diagnostic scope. Conditions are reported separately; deterministic proxy checks are not treated as human semantic-equivalence labels.

Condition	Eligible questions	Transformation and reporting rule
Query paraphrasing	all 1,019	deterministic wording change; original entities and fact constraints retained
Memory noise	all 1,019	three deterministic cross-conversation distractors; question and gold answer unchanged
Temporal reasoning	448 explicit temporal/date/order/duration questions	rewrite an existing temporal expression; no temporal requirement is added to an atemporal question
Multi-hop reasoning	216 original LoCoMo multi-hop questions	question and gold answer unchanged; only bridge distractors are perturbed
Inference and audit	seven conversations	2,000-replicate cluster intervals; 60 deterministic proxy items per included condition
Excluded analysis	not applicable	formula-extrapolated 64–4,096-item long-context scores are not included as performance evidence

Table S11 End-to-end supplementary-material profile on the same 159 questions. Calls/query and tokens/query are derived from run totals; token values are approximate because the recorded totals are rounded.

Method	Avg. (s)↓	P95 (s)↓	Calls/q↓	Tokens/q↓	$/query↓
BM25 RAG	8.2	16.5	1.0	2.70k	0.00041
Dense RAG	9.2	13.3	1.0	3.02k	0.00042
Hybrid RAG	11.2	23.0	1.0	3.27k	0.00046
MemoryBank	24.4	40.4	1.7	4.78k	0.00078
A-MEM	102.5	117.5	17.8	20.69k	0.00463
Mem0	53.4	63.4	14.1	19.69k	0.00375
Mem0g	179.1	193.1	77.3	54.03k	0.00959
Zep	92.3	111.8	33.5	35.85k	0.00676
AdaKVT-Mem	124.4	164.2	10.4	84.28k	0.01310

Table S12 Per-conversation overall F1 reconstructed from the current Primary per-question records. Each column is a held-out LoCoMo conversation; no conversation contributes questions to more than one split.

Method	conv-41	conv-47	conv-44	conv-49	conv-50	conv-42	conv-30
BM25 RAG	28.11	27.64	19.04	19.07	23.48	18.22	18.47
Dense RAG	33.52	32.95	22.26	24.15	27.82	22.39	20.58
Hybrid RAG	38.51	36.06	25.29	28.11	31.80	26.10	24.95
MemoryBank	35.72	34.46	24.62	26.46	29.79	24.16	22.96
A-MEM	44.59	40.42	31.36	33.67	37.94	31.48	33.84
Mem0	53.19	46.02	37.44	40.25	46.17	39.09	40.88
Mem0g	**56.42**	**50.02**	40.28	43.47	47.16	44.59	44.49
Zep	49.93	43.76	35.25	36.67	41.63	36.97	38.46
AdaKVT-Mem	49.08	48.61	**48.75**	**52.32**	**49.94**	**49.46**	**54.33**

Table S13 Complete conversation-cluster F1 comparison family. Bootstrap resamples complete conversations and reports question-weighted differences (20,000 replicates); exact sign-flip tests enumerate all 128 paired sign patterns. Holm values correct the eight displayed F1 contrasts. "Deletion" removes one test cluster without retraining or retuning.

Contrast	ΔF1	95% CI	W/T/L	Exact p	Holm p	Deletion range	Rank-1 deletions
vs. BM25 RAG	27.94	[24.11,31.62]	7/0/0	0.015625	0.125	[49.69,50.35]	7/7
vs. Dense RAG	23.53	[19.26,27.69]	7/0/0	0.015625	0.125	[49.69,50.35]	7/7
vs. Hybrid RAG	19.67	[15.24,23.92]	7/0/0	0.015625	0.125	[49.69,50.35]	7/7
vs. MemoryBank	21.51	[17.29,25.59]	7/0/0	0.015625	0.125	[49.69,50.35]	7/7
vs. A-MEM	13.83	[9.61,17.66]	7/0/0	0.015625	0.125	[49.69,50.35]	7/7
vs. Mem0	6.66	[2.08,10.63]	6/0/1	0.0625	0.125	[49.69,50.35]	7/7
vs. Mem0g	3.24	[−1.03,7.21]	5/0/2	0.15625	0.15625	[49.69,50.35]	7/7
vs. Zep	9.60	[5.20,13.61]	6/0/1	0.03125	0.125	[49.69,50.35]	7/7

Table S14 Strict budget-controlled LoCoMo mechanism comparison (Table S14). Every run contains 1,019 held-out questions from seven dialogue clusters; values are means over seeds 42–44. Δ is AdaKVT-Mem minus Mem0g after averaging seeds within question and then pairing whole conversations. The interval is a 20,000-replicate cluster bootstrap and p is the exact two-sided sign-flip result. The paired configuration validator permits only the memory backend to differ. For equal token, all 3,057 paired rows have identical logged full-input counts (620–696; mean 658.19); for equal call, every paired row has exactly three retrieval calls and the same two raw windows.

Protocol	Method	F1 mean ± SD	BLEU-1	Input tokens/q	Compliance	Paired result
Equal token (696)	AdaKVT-Mem	51.74 ± 0.04	43.93	658.19	100%	+0.35 [-3.39,3.70]; .859375
Equal token (696)	Mem0g	51.69 ± 4.20	43.09	658.19	100%	4/7 AdaKVT-Mem cluster wins
Equal call (3 retrievals)	AdaKVT-Mem	51.90 ± 1.57	44.87	949.02	100%	−2.88 [-6.59,0.37]; .218750
Equal call (3 retrievals)	Mem0g	55.09 ± 0.24	48.53	949.02	100%	2/7 AdaKVT-Mem cluster wins

Table S15 Strict-pair per-conversation F1 after averaging seeds 42–44 (Table S15). The final column is AdaKVT-Mem minus Mem0g. These are the seven paired cluster values used for the intervals and exact tests in Table S14; they expose the execution variation behind the equal-token Mem0g seed SD.

	Equal token (696)			Equal call (3 retrievals)		
Conversation	AdaKVT-Mem	Mem0g	Δ	AdaKVT-Mem	Mem0g	Δ
conv-30	54.16	49.62	+4.54	55.13	52.70	+2.43
conv-41	51.02	59.30	-8.28	50.92	62.87	-11.94
conv-42	52.11	49.71	+2.40	51.75	52.44	-0.69
conv-44	51.19	46.08	+5.10	50.83	51.41	-0.58
conv-47	50.03	55.10	-5.06	50.61	57.60	-6.99
conv-49	52.85	48.33	+4.52	53.65	52.22	+1.44
conv-50	51.65	52.38	-0.73	51.69	55.47	-3.78

Table S16 Task-layer update-semantics evaluation on 78 natural LongMemEval-S examples (Table S16). Relation and operation are task-subtype Macro-F1; transition and valid time are exact accuracies. Table S17 gives the fixed projection to the native AdaKVT contract. All non-Complete scores are percentages. Complete reports successful records/78 (thus 1/78=1.28%); 0/78 denotes an evaluated zero rather than an unavailable metric.

System	Candidate F1	Relation	Operation	Lifecycle	Valid time	Complete (n/78)
AdaKVT-Mem	62.91	78.06	78.06	98.72	70.51	1/78
w/o relation→operation mapping	62.91	78.06	30.67	98.72	70.51	0/78
w/o valid-time handling	64.24	78.06	78.06	98.72	6.41	0/78
w/o FSM enforcement	65.38	59.56	13.85	0.00	7.69	0/78
Mem0g matched	66.03	59.56	59.56	92.31	3.85	0/78
Raw context matched	16.03	59.56	59.56	92.31	0.00	0/78

Table S17 Fixed projection from natural-update task labels to the native AdaKVT relation–operation–state contract (Table S17). Task labels are measurement-layer subtypes, not additional lifecycle states or members of $\mathcal{O}_{rel}$. For AdaKVT-Mem, the resulting native relation Macro-F1 is 91.86 [82.15, 97.85], native operation Macro-F1 is 92.05 [81.48, 97.66], and transition accuracy is 98.72 [96.15, 100.00]. These are deterministic projections of Table S16 predictions, not independent trials.

Task subtype	Native relation	Native operation	Native state result
REPLACEMENT	update	revise	active → superseded; successor active
ACCUMULATION	refinement	update	active → superseded; successor active
REVERSAL	contradiction	revise	active → superseded; successor active
RELOCATION	temporal shift	shift	active → superseded; successor active
CORRECTION	contradiction	revise	active → superseded; successor active
NO_UPDATE	rejection	none	existing active cell retained

The legacy transition words `revised` and `invalidated` were removed from active labels because they describe update reasons, not lifecycle states. A traceable post-evaluation annotation and label-schema audit also normalizes two core-record value pairs (3→4 and $350,000→$400,000). Prediction-blind status was not contemporaneously recorded; therefore, an independent reviewer subsequently re-reviewed both changed value pairs from source evidence only, under the frozen guideline and without system predictions or prior annotations. Both were supported as high-confidence REPLACEMENT/REPLACE decisions, and every file in the blind-review package passed SHA-256 integrity verification. The package retains both snapshots: pre-amendment relation Macro-F1 77.51, lifecycle accuracy 92.31%, value-exact completeness 5/78, and complete success 0/78; current values are 78.06, 98.72%, 6/78, and 1/78, respectively. Complete success is a joint record-completion diagnostic, not a conditional transition-protocol success measure. Changed-field hashes, the independent-review protocol and outcome, regenerated metrics, and the native-contract result file are retained in the complete archive.

Table S18 Additional write-path coverage and controlled safety diagnostics (Table S18). Natural-update scores use the 78-example double-annotated set and 95% bootstrap intervals. The 54 augmentations and 660 hard negatives are constructed diagnostics and are not pooled with the natural set.

Split	Metric	Score (%)	95% interval / scope
Natural ($n = 78$)	predecessor Recall@1	36.11	[23.94,47.22]
Natural ($n = 78$)	predecessor Recall@5	88.89	[80.28,95.89]
Natural ($n = 78$)	predecessor Recall@10	94.44	[88.73,98.65]
Natural ($n = 78$)	predecessor Recall@20	95.83	[90.54,100.00]
Natural ($n = 78$)	value-exact completeness	7.69	[2.56,14.10]
Natural ($n = 78$)	complete-update success	1.28	[0.00,3.85]
Controlled ($n = 54$)	candidate extraction F1	100.00	constructed augmentation diagnostic
Hard negatives ($n = 660$)	joint reject/no-op accuracy	100.00	five constructed conditions, 132 each
Hard negatives ($n = 660$)	destructive false-update rate	0.00	five constructed conditions, 132 each

Table S19 Aggregate task-relation class support and outcomes for AdaKVT-Mem (Table S19). These precision/recall/F1 values are calculated from the released 78-example aggregate confusion matrix; Table S16 reports the protocol's partition-pooled Macro-F1.

Task subtype	Support	Precision (%)	Recall (%)	F1 (%)
ACCUMULATION	29	90.00	93.10	91.53
CORRECTION	1	0.00	0.00	0.00
NO_UPDATE	6	100.00	83.33	90.91
RELOCATION	6	85.71	100.00	92.31
REPLACEMENT	31	90.32	90.32	90.32
REVERSAL	5	100.00	100.00	100.00

The singleton CORRECTION class makes a reliable class-level estimate impossible; this imbalance is a limitation rather than evidence of zero capability.

The released per-sample record also supplies a fixed-order, mutually exclusive completion account: 48 candidate-set misses, 12 rank-1 predecessor misses, 17 strict before/after-value-pair misses after all other recorded checks, and one complete record. Conditional on exact candidate reconstruction and a rank-1 predecessor, all 18 retained records have correct relation, operation, lifecycle, valid time, and provenance; 1/18 has an exact value pair.

Table S20 Configuration-level LoCoMo engineering comparison on 1,019 questions from seven held-out conversations (Table S20). Category-macro weights categories equally; question-weighted follows the 198/216/546/59 question counts. These are complete-system operating points with native construction and internal retrieval; the strictly paired two-backend comparison is reported separately in Table S14.

Method	Single	Multi	Temporal	Open-form	Macro	Weighted
BM25 RAG	23.68	10.49	25.94	24.57	21.17	22.15
Dense RAG	29.50	14.47	30.05	28.67	25.67	26.56
Hybrid RAG	32.59	18.70	34.17	31.34	29.20	30.42
MemoryBank	31.26	17.81	31.85	28.82	27.44	28.58
A-MEM	36.66	25.29	40.76	33.41	34.03	36.26
Mem0	40.79	33.93	47.72	47.43	42.47	43.43
Mem0g	39.70	33.30	54.10	53.43	45.13	46.85
Zep	38.57	29.79	43.92	54.31	41.65	40.49
AdaKVT-Mem	50.36	48.42	50.60	50.64	50.00	50.09

Table S21 Equal-token seed sensitivity from the archived 1,019-question records (Table S21). Δ is the per-question mean AdaKVT-Mem-minus-Mem0g F1 within the stated seed; the leave-one-seed-out row averages the other two Δ values. Within each seed, every paired backend row has equal input-token count. Candidate identity/rank checks operate on logged retrieval records. Missing serialized evidence-payload hashes and provider response identifiers locate the variation but do not attribute it between payload construction, provider execution, or an external model revision.

Seed	AdaKVT-Mem F1	Mem0g F1	Δ F1	Mem0g mean input tokens
42	51.79	46.85	+4.93	657.92
43	51.72	53.87	-2.16	658.56
44	51.70	54.35	-2.65	658.10
Omit 42	–	–	-2.40	mean of 43, 44
Omit 43	–	–	+1.14	mean of 42, 44
Omit 44	–	–	+1.39	mean of 42, 43

Seed pair	Candidate-set Jaccard	Ranked list identical	Input tokens identical	Answer identical	F1 correlation
42 vs. 43	1.000	100.00%	1.37%	5.40%	0.927
42 vs. 44	1.000	100.00%	1.28%	4.81%	0.930
43 vs. 44	1.000	100.00%	1.28%	7.85%	0.947