TACTICO

Momentum (L3.1)

System Card

Independent Data · 2026-11-22

Contents

1. Abstract and key findings **2**
1.1 Findings . . . 2

2. System and treatment selection **3**
2.1 Memory path under test . . . 3
2.2 Promotion lock . . . 3

3. Evaluation design **4**
3.1 Frozen cohorts . . . 4
3.2 VibeMemBench fixture qualification . . . 4
3.3 Common budgets and matched readers . . . 5
3.4 Assignment and trial units . . . 5

4. Scoring and statistical methods **6**
4.1 Retrieval . . . 6
4.2 Conversational answers and V2 . . . 6
4.3 Executable coding . . . 6

5. Primary results **7**
5.1 Retrieval on the broader holdout . . . 7
5.2 Conversational answer accuracy . . . 7
5.3 Compact V2 and matched-reference pair outcomes . . . 7
5.4 VibeMemBench strict task resolution . . . 8

6. Exploratory results and mechanism diagnostics **9**
6.1 Short-turn reranking . . . 9
6.2 Labeled source coverage and usable excerpts . . . 9
6.3 Typed claims and guard checks . . . 9
6.4 Timing and indexing costs . . . 10

7. Infrastructure, actors, and isolation **10**
7.1 Boat allocation . . . 10
7.2 Model actors and activity counts . . . 10
7.3 Clean-room boundary . . . 11

8. Independent benchmark context **11**
8.1 Matched and separately reported results . . . 11
8.2 All four L3.1 readers with external context . . . 12
8.3 Sol-only view . . . 12
8.4 Five external sources and their protocols . . . 13
8.5 Operational deviations and interpretation limits . . . 13

9. Reproduction and artifact provenance **14**
9.1 Audit trail . . . 14

Appendix A. Historical controls and ablations **15**

References **16**

1. Abstract and key findings

Momentum (L3.1) adds source-first semantic retrieval to Ledger. This card presents the finalized clean-room study and adds five external reports to place its scores in context. L3.1 has 780 scorable results from 800 assignments across LoCoMo, LongMemEval-S, compact LongMemEval-V2 Small, and VibeMemBench. Including historical controls, the archive contains 2,337 scorable outcomes from 2,400 Momentum assignments and all 400 matched reference judgments. The external figures retain their original evaluation protocols.

1.1 Findings

- **L3.1 delivered higher LoCoMo and LongMemEval-S recall than its episodic predecessor.** Its macro delivered Recall@5 was 50.85% and 90.05%, respectively. It trailed matched Aletheia hybrid by 5.24 percentage points on LoCoMo and led matched AgentMemory BM25 by 3.40 points on LongMemEval-S.
- **End-to-end answers trailed the executed references.** L3.1 answered 84/200 LoCoMo and 94/200 LongMemEval-S reader trials correctly, alongside reference results of 95/200 and 99/200. Those pooled counts summarize reader judgments on shared questions.
- **A separate short-turn reranker reached 61.88% LoCoMo Recall@5.** The research arm reordered the same candidate source sets as L3.1. It was examined separately after the universal reranker missed its development gate. Downstream answering was outside that experiment.
- **Compact V2 answer accuracy was 10.5%.** L3.1 answered 21/200 questions correctly under the official deterministic scorer on the fixed text projection of Small.
- **Coding results varied by reader and harness.** L3.1 resolved 60/180 comparable VibeMemBench trials across 200 assignments. Sol resolved 33/50 and Terra 26/50 in the repository-tool harness. Patch-only Qwen resolved 0/50 and patch-only Grok 1/30, with 20 Grok transport deadlines outside the comparable denominator.

L3.1 is the development-selected `l3-semantic-source-first` build. The matched comparisons use shared inputs and scoring. Section 8 shows the separately published studies beside Tactico's L3.1 results. Historical controls appear in Appendix A.

2. System and treatment selection

2.1 Memory path under test

Ledger links each admitted excerpt to its original evidence event by an immutable identity. It applies scope, provenance, time, conflict, supersession, and forgetting rules before ranking. L3.1 ranks original sources, selects one eligible passage per source, and constructs the reader's evidence packet.

The public pilot controlled source projection, cutoff and budget by supplying both L2 and L3 with the same episodes and zero typed claim proposals. The separate synthetic study exercised typed claims. In the six labeled LoCoMo pilot cases, two gold turns moved past the five-source cutoff, producing the 36.1% to 13.9% Recall@5 decline that motivated this study.

Table 1. Mechanisms compared during development

Treatment	Mechanism
Native L2 and prior L3	Existing native episode ranking and filtering on L2, and the original L3 episodic BM25 path.
L2-style replay inside L3	Native-style scoring and deduplication across the episodes eligible for L3.
L3 lexical reranking	Rerank lexical candidates using rare terms, phrase coverage, and source diversity.
Narrow claims	Ground a restricted set of source-attributed relations before queries and route qualifying requests through Ledger typed facts.
Ordinary semantic fusion	Combine sparse BM25 and BGE-small dense rankings within the same authorized history.
Source-first semantic	Collapse retrieval signals to original source IDs. Fuse up to 20 distinct sources per channel using reciprocal rank fusion at k=20. Choose one eligible passage per selected source.
Short-turn cross-encoder	Reorder the same candidate passages with pinned MiniLM and the base ranking. Cross-encoder and base reciprocal-rank weights use 2:1.

2.2 Promotion lock

The study selected source-first on development results and froze it before accessing holdout scores. A universal MiniLM challenger needed a mean gain of at least one point across the two development Recall@5 scores, with neither declining by more than one point. LongMemEval-S fell by 5.25 points. Researchers later froze the exploratory turn-only LoCoMo test after seeing primary holdout results and before opening reranker outcomes.

3. Evaluation design

3.1 Frozen cohorts

Table 2. Cohorts and evaluation units

Benchmark	Calibration	Broader retrieval holdout	End-to-end IDs per reader and arm
LoCoMo	653 questions in 3 conversations · 502 labeled	1,297 questions in 7 conversations · 1,004 labeled	50 questions, 10 from each of 5 categories
LongMemEval-S	24 questions · 20 labeled	120 disjoint questions · 103 labeled	50 questions across six substantive types and abstention
LongMemEval-V2 Small	Text projection and scorer smoke checks	Official evidence labels unavailable	50 questions, 25 web and 25 enterprise
VibeMemBench	75 ordinary coding targets in ordered shortlist	Earlier-history evidence, without gold retrieval labels	50 qualified tasks across 45 repositories and 7 languages

LoCoMo split whole conversations, favoring prior-exposed conversations for development. LongMemEval-S used disjoint histories in the cleaned release and sampled by category, with seven reader cases per stratum except eight preference cases. Source IDs and normalized text excluded earlier pilot questions (24 LoCoMo, 28 LongMemEval-S). Gold evidence units are dialogue turns and sessions, respectively. LoCoMo · Cleaned LongMemEval-S

V2 excluded 36 earlier IDs, leaving 258 eligible deterministic text cases. Each domain includes nine dynamic, eight procedure, and eight static questions, with a 100-trajectory Small haystack. The compact projection keeps up to 12,000 bytes of actions, thoughts, and first appearances of accessibility text per trajectory. A separate uncapped text projection uses the same IDs. Both omit screenshots. Public V2 omits gold trajectory IDs. V2 dataset · Published V2 results

3.2 VibeMemBench fixture qualification

Qualification required a pinned base and image, an admissible earlier same-repository trajectory, a passing gold control, and an unchanged base that failed a target check yet preserved regression checks. Fifty qualifying ordinary tasks were frozen through shortlist position 63 with 13 control failures. Twelve entries lay beyond the qualified prefix. Solvers received the shared issue and pristine base. The separate evaluator retained the gold patch and private tests. VibeMemBench dataset · Paper

3.3 Common budgets and matched readers

QA uses at most five distinct sources, 8,000 rendered UTF-8 evidence bytes and 20 internal candidates. Chunks share one original source ID. Coding uses five episodes, a 1,200-token proxy and at most 920 bytes per episode, replayed in earlier pull-request order with zero claim proposals. A 10,000-byte native query cap affected one coding retrieval query. Every task prompt retained the full public issue.

Table 3. Reader runtime and interfaces

Reader	Conversational QA and V2	VibeMemBench
GPT-6.2 Sol	Cold Codex subagents. Up to four assigned QA questions or up to two V2 questions per context, with narrow packet and answer paths.	Fresh task-level Codex context with checkout tools and public tests.
GPT-6.1 Terra	Same Codex transport, blinded memory arm codes, and independent guest assignments as Sol.	Fresh task-level Codex context with checkout tools and public tests.
Qwen3.5-9B	Pinned quantized checkpoint under llama.cpp, isolated provider requests on assigned Boat guests.	Task and bounded public source context supplied in a patch-only harness on Boat.
Grok 4.7 Build	Routed through native Build CLI sessions from the controller with the assigned Boat workspace.	Native Build, bounded packet-to-patch format, and verified application on the assigned Boat worktree.

Matched plans share IDs, reader settings, prompt template, evidence cap, and judge. Each reader sees opaque case and source IDs. Byte hashes bind baseline, L3.1, and matched-reference packet plans. Later comparisons reuse accepted sealed answers.

The Qwen checkpoint was `Qwen3.5-9B-UD-Q4_K_XL` under llama.cpp b12580. Conversational QA capped output at 128 tokens. The coding wrapper requested a 16,384-token context, a 1,300-token output cap, and temperature zero. Grok used native Build CLI 1.0.77 and Low effort in the attested lanes. Sol and Terra used managed Codex model IDs with their role-specific bounded tool interface.

3.4 Assignment and trial units

One end-to-end trial is one case, reader, and memory arm. QA contributes 1,200 primary assignments, V2 and VibeMemBench 600 each, and matched references 400 additional QA trials. A VM can host multiple cold contexts and a context several QA trials. Retrieval observations are indexed by query and arm. Paired analyses use shared scorable cases and retain separate outcome and infrastructure statuses.

4. Scoring and statistical methods

4.1 Retrieval

Recall@K divides distinct labeled sources found in the top K by all sources labeled for that question and macro-averages across questions. Packet Recall@5 uses the budgeted rendered shortlist. Candidate Recall@5/10/20, any-gold, and all-gold measure earlier coverage. The primary archive has 12,753 rows across 1,417 queries and nine arms. The exploratory table includes 1,297 new reranker observations among 5,188 rows.

Retrieval intervals use 10,000 paired scope-cluster draws under seed 20261122. The resampling unit is one of seven LoCoMo conversations or 103 labeled LongMemEval-S scopes. Few LoCoMo clusters limit reliability beyond the frozen set.

4.2 Conversational answers and V2

One pinned Qwen judge scores all LoCoMo and LongMemEval-S arms. Bare abstention on an answerable question scores incorrect. Export raw and strict verdicts and LoCoMo answerable token-overlap F1. Four strict overrides concern one Qwen LongMemEval-S case across all four arms. Pooled answer intervals were added after scoring and cluster all reader judgments for each sampled scope.

V2 isolates the boxed final answer and applies the released deterministic function for the selected question type. An unknown answer scores false on the selected subset. The primary V2 denominator includes every completed scored answer, with results by model, arm, domain and category. The benchmark input represents text-only Small, so public full-multimodal headline results serve as protocol context.

4.3 Executable coding

An official task resolves only if every declared fail-to-pass and pass-to-pass check passes. The grader uses the pinned evaluator and image, with networking disabled for the named grading containers. The main strict submission uses the original candidate response bytes. The declared official-test overlay removes collisions confined to enumerated test paths while keeping implementation hunks. Raw, overlay and candidate application outcomes are retained, including no applicable patch and setup errors.

A separate terminal-newline sensitivity appends exactly one line-feed byte to eligible original response text and reruns the declared patch validation and grading. It reuses responses and is separately reported. Verified setup failures caused by candidate code count as unresolved in the reconciled view. Infrastructure exclusions stay unscored. The strict result table gives resolved/scorable counts and all cells list 50 assigned tasks.

5. Primary results

5.1 Retrieval on the broader holdout

Table 4. L3.1 delivered macro Recall@5 at the common budget

Dataset	Labeled	Momentum (L3.1)	Matched retrieval reference
LoCoMo	1,004	50.85%	Aletheia hybrid 56.09%
LongMemEval-S	103	90.05%	AgentMemory BM25 86.65%

Against the matched retrieval components, L3.1 was 5.24 points lower than Aletheia hybrid on LoCoMo (exploratory 95% interval -6.00 to -4.31), and 3.40 points higher than AgentMemory BM25 on LongMemEval-S (0.00 to 7.44). The latter lower endpoint rounds to zero. Paired changes from prior Momentum builds appear in Appendix A.

5.2 Conversational answer accuracy

Table 5. L3.1 and matched QA reference. Correct out of 50 per reader, pooled out of 200.

Dataset / reader	Momentum (L3.1)	Matched reference
LoCoMo / GPT-6.2 Sol	22	27
LoCoMo / GPT-6.1 Terra	23	23
LoCoMo / Qwen3.5-9B	14	18
LoCoMo / Grok 4.7 Build	25	27
LoCoMo / pooled	**84**	**95**
LongMemEval-S / GPT-6.2 Sol	28	28
LongMemEval-S / GPT-6.1 Terra	26	27
LongMemEval-S / Qwen3.5-9B	14	17
LongMemEval-S / Grok 4.7 Build	26	27
LongMemEval-S / pooled	**94**	**99**

L3.1 and the matched references differed by -5.5 points on LoCoMo and -2.5 on LongMemEval-S when pooling reader judgments. Equal totals can arise from different correct cases. On LongMemEval-S, Sol scored 28 with both L3.1 and AgentMemory, with three questions correct only in each arm.

5.3 Compact V2 and matched-reference pair outcomes

Table 6. Compact V2 official answer accuracy and paired QA references

Reader	V2 L3.1 correct / 50	LoCoMo L3.1 only / reference only / tied	LongMemEval-S L3.1 only / reference only / tied
GPT-6.2 Sol	7	3 / 8 / 39	3 / 3 / 44

Reader	V2 L3.1 correct / 50	LoCoMo L3.1 only / reference only / tied	LongMemEval-S L3.1 only / reference only / tied
GPT-6.1 Terra	6	4 / 4 / 42	3 / 4 / 43
Qwen3.5-9B	2	0 / 4 / 46	0 / 3 / 47
Grok 4.7 Build	6	2 / 4 / 44	2 / 3 / 45

L3.1 answered 21/200 compact V2 trials correctly and abstained on 145/200. A separate full-text comparison reports evidence differences on the same 50 questions. That comparison measured retrieval and packet contents only.

Supplemental 95% intervals use 2,000 paired scope-cluster draws. L3.1 minus the executed references was -5.5 points on LoCoMo (-13.04 to 1.79) and -2.5 points on LongMemEval-S (-11.0 to 4.5). The paired counts are the recorded results. The intervals are exploratory and the study declared no equivalence margin.

5.4 VibeMemBench strict task resolution

Table 7. L3.1 strict task resolution and grader coverage

Reader	Resolved / comparable	Official scored / assigned
GPT-6.2 Sol	33/50	50/50
GPT-6.1 Terra	26/50	49/50
Qwen3.5-9B	0/50	50/50
Grok 4.7 Build	1/30	30/50

L3.1 resolved 60/180 comparable tasks across 200 assignments. Grok's 20 unscored events were native provider deadlines. Its 30 finalized replies yielded three Git-applicable patches, 20 rejected patches, and seven no-patch responses. The Qwen strict wrapper accepted no patches. The one-line-feed sensitivity resolved 2/7 eligible Grok outputs and 0/42 eligible Qwen outputs. These sensitivity counts are separate from strict results.

The official grader left one L3.1 Terra result unscored because candidate compilation failed. The comparable column counts that verified candidate failure as unresolved. It retains infrastructure-only events outside the scoring denominator. Each coverage fraction retains the original denominator of 50 assigned tasks.

6. Exploratory results and mechanism diagnostics

6.1 Short-turn reranking

The separate LoCoMo MiniLM track reached 61.88% delivered Recall@5 on the 1,004 labeled questions. The paired difference was +11.03 points against L3.1 (exploratory 95% interval 9.42 to 12.66) and +5.79 against matched Aletheia hybrid (4.86 to 6.85). The candidate source sets matched L3.1 exactly for all 1,297 queries, and candidate Recall@20 stayed 68.72%. Reordering raised complete-evidence@5 from 46.71% to 56.87%. The study reused the existing LoCoMo holdout after viewing the primary aggregate and froze the short-turn policy before inspecting this reranker's outcomes. This track measures retrieval on short passages, with no downstream answering trial.

6.2 Labeled source coverage and usable excerpts

On the exact answer-reader sample, packet Recall@5 went from 39.44% to 46.66% on the 40 labeled LoCoMo cases and from 80.23% to 87.60% on the 43 labeled LongMemEval-S cases, relative to prior L3. A separate literal-span diagnostic counted how often the packet included the token-normalized reference phrase. L3.1 delivered at least one labeled source on 23/40 and 39/43 answerable questions, while exact phrase matches appeared on 10/40 and 17/43. Prior L3 values were 18/40 and 38/43 for source presence, and 10/40 and 20/43 for exact phrase presence. Source recall, phrase coverage, and graded correctness measure different things. A phrase check can miss equivalent wording.

An exploratory same-budget passage packer tried adding nearby windows from the already selected sources on 24 LongMemEval-S development questions. It preserved the primary excerpts and source order. On 20 answerable questions, literal phrase coverage rose from 8/20 to 9/20 and labeled-source coverage stayed at 19/20. Thirty-five passages were added across 23 cases and source byte checks passed all 24. The experiment did not generate answers.

6.3 Typed claims and guard checks

The isolated typed-claim path checked provenance, scope, time cutoffs, supersession, and forgetting. Its strict candidate filter found one eligible span across 5,882 LoCoMo turns and none across 6,913 flattened LongMemEval-S sessions. Extraction yielded zero accepted LoCoMo claims. A broader role-aware LongMemEval-S prototype admitted two claims in its 24-case development trial and routed no queries to them. It tied L2-style replay at 83.0% packet Recall@5 across 20 labeled cases. Observation reranking changed source membership on two cases. Narrow claim variants also preserved the exact candidate and packet IDs of their paired ranker controls in the 1,417-query holdout. These results identify low claim coverage in the tested replay adapters.

6.4 Timing and indexing costs

In 24 paired LongMemEval-S development cases on one 8-vCPU Boat guest, ranked packets were identical between empty and warmed document-vector caches. Cold median/p95 retrieval was 10.686/12.944 seconds. Warm retrieval was 28.682/35.174 ms with the query freshly encoded on both paths. A separate document-index timer measured 10.655 seconds median. The backend timer excludes reader generation, final packet rendering, and provider transport. A separate preindex prototype passed scope, cutoff, and forgetting checks. It was outside the frozen benchmark.

7. Infrastructure, actors, and isolation

7.1 Boat allocation

Table 8. Distinct private Boat guests by registered study role

Role	Guests
Conversational QA readers including matched references	32
Compact V2 readers	24
VibeMemBench reader hosts	21
Retrieval workers	17
Graders and fixture hosts	11
Question-free claim extractor	1
Total	**106**

All 106 registrations requested the large VM shape and an empty inherited environment. Provider receipts confirm 8 vCPU and 16 GB on all 106. The final provider inventory verified 106 of 106 archived. The finalized resource census is dated 12:09:04 UTC on November 22, 2026. One API sample observed 59 active guests, and receipt lifetimes overlapped at most 60, including lifecycle overhead. Two resumes reused one retrieval VM. The earlier pilot had a separate 23-VM fleet. This census covers the finalized study. External reports supply published scores without adding Boat runs to the census.

7.2 Model actors and activity counts

The agent census at 11:52:36 UTC identified 47 distinct coordination, engineering, dispatch, audit, and reporting subagents plus the root coordinator. It covers the expanded study beginning at 09:55:10 UTC on November 22, 2026, and includes reporting roles active at that snapshot. Cold evaluated readers were counted separately by their lane ancestry. A team role can create many evaluated contexts over the run.

Table 9. Cold evaluated Codex contexts

Workload	Intended contexts	Spawned threads
Conversational QA including matched references	264	265
Compact V2	156	157
VibeMemBench native coding	300	311
Total	**720**	**733**

For Sol and Terra, 264 accepted conversational groups cover 800 QA answers, and 156 V2 groups cover 300 answers. Each native coding context targets one task. Spawn counts include pre-read and replacement threads. The separate native-runtime census verified 400 conversational QA and 150 V2 case-level executions for each of Qwen and Grok. It also verified 150 Qwen coding attempts. The 150 Grok coding assignments produced 92 finalized replies and 58 transport-only outcomes across the three arms. A separate provider sample observed six simultaneous Grok CLI processes, with four coding and two QA.

7.3 Clean-room boundary

Source preparation, rankers, readers, and graders have role-specific access. Retrieval workers receive authorized histories and questions under opaque identifiers. The broader claim experiment was staged with records only and never received question or grade files. Private Boat readers receive their assigned prompts or pristine worktrees, a bounded outbox, and public test access where permitted. Graders retain gold answers, label mappings, reference patches, and hidden tests. Reader-to-grader transfer binds the case assignment and exact response bytes by hash. Gold joins, result classification, and redaction happen on the grading side.

Cold Codex threads run with minimal inherited context. Qwen inference stays on assigned guests. The Grok controller invokes managed provider sessions locally and uses their assigned Boat workspaces and bounded outputs. The process boundary constrains benchmark materials, tools, and history exposure for the study. Prior training exposure and the external provider infrastructure are outside what the study can attest.

8. Independent benchmark context

8.1 Matched and separately reported results

Section 5 uses the same histories, four readers, budget, and judge for the matched Aletheia and AgentMemory runs. The five external reports keep their own protocols. Mnemoverse audits its own saved answers. Aletheia benchmark · AgentMemory benchmark receipts

8.2 All four L3.1 readers with external context

Table 10. L3.1 and separately published answer accuracy or task resolution. Parentheses show L3.1 successes out of scorable cases.

Benchmark	L3.1 Sol	L3.1 Terra	L3.1 Qwen	L3.1 Grok	Mem0	Supermemory	Zep / Graphiti
LoCoMo	44.0% (22/50)	46.0% (23/50)	28.0% (14/50)	50.0% (25/50)	77.68% [3]	73.53% [3]	63.83% [3]
LongMemEval-S	**56.0%** (28/50)	52.0% (26/50)	28.0% (14/50)	52.0% (26/50)	**56.00%** [3]	66.07% [3]	79.80% [3]
LongMemEval-V2	14.0% (7/50)	12.0% (6/50)	4.0% (2/50)	12.0% (6/50)	-	-	-
VibeMemBench	**66.0%** (33/50)	52.0% (26/50)	0.0% (0/50)	3.3% (1/30)	62.8% [4]	-	-

Each L3.1 reader had 50 assignments per benchmark. Grok had 30 scorable coding attempts after 20 provider deadlines. Sol and Terra used repository tools. Qwen and Grok used patch-only wrappers. Gray identifies the highest displayed OmniMemEval value per benchmark. A dash marks an unfilled comparison. OmniMemEval · VibeMemBench

8.3 Sol-only view

Table 11. L3.1 with GPT-6.2 Sol and selected external reports under their original protocols.

Benchmark	L3.1 with Sol	External score	Protocol difference
LoCoMo	44.0% (22/50)	77.68% Mem0 [3]	OmniMemEval uses a different reader, judge, and LoCoMo cohort.
LongMemEval-S	**56.0%** (28/50)	**56.00% Mem0** [3]	Equal displayed accuracy across Tactico's 50 cases and OmniMemEval's 500-case benchmark.
LongMemEval-V2	14.0% (7/50)	69.9% Codex and 74.9% AgentRunbook-C [5]	Compact text-only Small alongside the published full multimodal Small runs.
VibeMemBench	**66.0%** (33/50)	62.8% Mem0 with DeepSeek V4 Pro [4]	Different solvers, task cohorts, and coding harnesses.

In both tables, light cyan marks equal displayed LongMemEval-S values. Statistical equivalence would require a matched protocol and a declared margin. Stronger cyan marks Sol's 66.0% VibeMemBench score, 3.2 percentage points above the separately published Mem0 run. A matched Sol memory-off test would quantify the contribution of L3.1 to its coding result. OmniMemEval · VibeMemBench · LongMemEval-V2 Table 2

8.4 Five external sources and their protocols

- **[1] Mnemoverse rubric audit.** GPT-5 rescored the same 152 saved non-adversarial answers from LoCoMo conversation 26. The run had used a GPT-5 reader and top-200 retrieval. The score moved from 90.13% with the lenient rubric to 48.03% with the stricter rubric, with 64 verdicts changing. The stricter criteria allowed less leeway for partial answers and approximate dates. Mnemoverse evaluation report
- **[2] Separate Mem0 top-20 pipeline.** The authors report 57.68% on LoCoMo and 49.00% on LongMemEval, with nominal dataset sizes of 1,986 and 500 questions. The paper names no LongMemEval edition. GPT-5-nano extracted facts. GPT-5-mini answered from the top 20 vector-retrieved facts. The assessment used a GPT-5-mini judge and three-vote majority. Study Table 3
- **[3] OmniMemEval.** GPT-4.1-mini answered and GPT-4o-mini judged 1,540 non-adversarial LoCoMo questions and 500 cleaned LongMemEval-S questions. Reported context differed by backend. Its LoCoMo Zep and LongMemEval-S Graphiti values use separate cloud and self-hosted deployments, respectively. On LongMemEval-S, mean measured context included 856 tokens for Mem0 and 117,106 for self-hosted Graphiti. The percentages are reproduced as reported. OmniMemEval primary results
- **[4] VibeMemBench paper control.** The published DeepSeek V4 Pro experiment used 111 tasks with four seeds in mini-swe-agent. The one-record Mem0 run scored 62.8% and the same solver scored 67.1% with memory off. The difference was -4.3 percentage points, with paired interval -9.68 to +1.35 points. Eleven of the paper's 12 memory-system and solver pairings scored at or below their corresponding baseline. VibeMemBench Tables 1 and 2
- **[5] Official LongMemEval-V2 result.** Full multimodal Small reports 69.9% for Codex and 74.9% for AgentRunbook-C. The comparison used a GPT-5.4-mini controller at xhigh effort and the same fixed Qwen3.5-9B downstream reader. L3.1 used the compact text-only deterministic subset. LongMemEval-V2 Table 2

8.5 Operational deviations and interpretation limits

Four earlier Terra Vibe cases had launcher quoting failures without model-produced patches and are excluded only from the comparable denominator. An additional Terra action was blocked by automatic approval review for suspected scope drift. It was retained as a blocked outcome and left the checkout pristine. Verified candidate compilation or parsing failures are assessed as unresolved, even when tests could not start. Grok provider timeouts and transport failures remain distinguishable from applicable patches that failed official tests. The Qwen coding format gave zero strict resolutions in every arm. The one-line-feed sensitivity recovered some Git-applicable outputs, with official results reported separately.

The held-out answering sample is 50 questions per benchmark and same-question judgments across readers are dependent. Frozen Vibe

qualification narrows the task population to ordinary reproducible repository changes. The compact V2 cohort excludes screenshots and question types that require a separate judge. Tool-using Codex coding and patch-only coding answer different interface questions. Published vendor percentages inform context. The principal evidence for comparison comes from matched cohort, model, budget, and scorer.

9. Reproduction and artifact provenance

9.1 Audit trail

The reproducible sequence is to verify source bytes, recreate or load the frozen cohort and opaque case assignment, stage only assigned public materials on workers, reproduce the frozen ranker and prompt, preserve the returned bytes and runtime attribution, check content-addressed transport, and grade each exact key on its authorized grader. Hidden gold and raw model outputs stay on the assigned private side. Recompute aggregate and pair summaries from the text-free CSV/JSONL exports. Because opaque IDs were created per freeze, exact agreement of identifiers uses the retained cohort artifacts and sealed mapping.

Table 12. Primary cohort and system pins

Artifact	Revision or SHA-256
LoCoMo source	`3eb6f2c585f5e1699204e3c3bdf7adc5c28cb376`
LoCoMo dataset SHA-256	`79fa87e90f04081343b8c8debecb80a9a6842b76a7aa537dc9fdf651ea698ff4`
LongMemEval-S cleaned release	`98d7416c24c778c2fee6e6f3006e7a073259d48f`
LongMemEval-S dataset SHA-256	`d6f21ea9d60a0d56f34a05b609c79c88a451d2ae03597821ea3d5a9678c3a442`
V2 dataset and official scorer revisions	`f152293e235517d504809563c833d7190b8c713b`, `2cc8c540bdb87fe6761629b585e727e1c4704520`
V2 official scorer SHA-256	`f8ba659e237c4282a07a92538090ea500ce8483820b390b1d19af187c01417ac`
VibeMemBench dataset and official evaluator revisions	`11164976b4f920cdb0d0cd570ca572e49acea1e8`, `faf754671f5b850e1b4d1ff870db37a6a2480df6`
VibeMemBench selected cohort SHA-256	`e39f9abf5e9795192fe6ebd7cd1a26a0951f74c034d41413e3c3022e1a6e3b04`
VibeMemBench strict three-arm receipt SHA-256	`ba354135543f891640b23cf19b6f7ff5c20c6650f7b0c0287e4c2aa14b3ad22c`
L3.1 code archive SHA-256	`b2af70b0ff9af7478094c6a1df89acc3c49ae5dff8a10ad682d492e2751c8b4b`
Primary promotion lock SHA-256	`a5c1458b13e6e9f8aceeee7a4202651e3bfee03b1d4b017a8a1bd910d963da86`
Exploratory short-turn lock SHA-256	`1e3058a362c58a8bcebe458a8b85bdab56bf20ad6d64e7dcd0d32f8c8e7e0f7f`
Pinned Qwen judge checkpoint SHA-256	`6f5d30666c2d8ae16a306e616d95341dcf3cc46810df84d7e6f5a7d1e4c1b293`
Aletheia and AgentMemory component revisions	`e6145b3b3b06ee1ec3ad0f116f0a5274ff3d9520`, `3f8784d10c2e996873d4392b85226a5d042def0e`
BGE-small ONNX model revision	`ea104dacec62c0de699686887e3f920caeb4f3e3`
Expanded retrieval data ZIP SHA-256	`be95a091cf3932863ae55c44b19ea5c6180e76129804c09a89fdb34dbd112667`

Artifact	Revision or SHA-256
L3.1 benchmark data ZIP SHA-256	`39ccdfc53a755361f03bda4f53931172eb02dd18bb5859670bd256fa38814b2c`
Final resource census SHA-256	`e22addf1c1c4d2179aa847a1e993de689e2e67a6ce0ae860b9d05b23c289290f`

The checked retrieval archive contains 17 member checksums. The paired answer export carries original graders' statuses, row-level assignment keys, strict model outcomes, benchmark and model totals, and paired comparator counts. The release manifest lists the companion PDF, plots, tables and data checksums. Resource records describe VM lifecycle and agent orchestration separately from scientific denominators.

Appendix A. Historical controls and ablations

Table 13. Development ablations. Delivered macro Recall@5.

Treatment	LoCoMo n=502 labeled	LongMemEval-S n=20 labeled
Native L2 and L2-style replay	48.71%	83.00%
Prior L3	47.94%	83.00%
Lexical reranking, with or without narrow claims	53.03%	83.00%
Ordinary semantic, with or without narrow claims	56.32%	89.25%
L3.1 source-first semantic	56.65%	91.75%
Matched open component	60.95%	88.25%
MiniLM challenger	66.06%	86.50%

On held-out LoCoMo, native L2 scored 43.89% and prior L3 scored 42.51% delivered Recall@5. L3.1's gain over prior L3 was 8.34 points (95% scope-cluster interval 6.95 to 10.00), with 158 increases, 47 decreases, and 799 ties across labeled cases. On LongMemEval-S, the two controls each scored 85.44%. L3.1 gained 4.61 points (1.70 to 8.09), with 11 increases, one decrease, and 91 ties. L2 and L2-style replay returned the exact same ordered candidate and packet IDs on all 1,417 holdout queries.

Table 14. Historical and L3.1 answer controls. Counts pooled across readers.

Benchmark	Native L2	Prior L3	L3.1
LoCoMo	86/200	81/200	84/200
LongMemEval-S	103/200	102/200	94/200
Compact LongMemEval-V2	21/200	21/200	21/200

Relative to prior L3, the supplemental 95% pooled answer interval was +1.5 points on LoCoMo (-5.09 to 10.12) and -4.0 points on LongMemEval-S (-12.5 to 3.0). LoCoMo L3.1 vs L2 was -1.0 point (-7.55 to 5.95). LongMemEval-S L3.1 vs L2

was -4.5 points (-13.0 to 3.0). These exploratory intervals resample whole history scopes, preserving all model-question pairs in a selected scope.

Table 15. Historical coding controls. Official resolved/scorable before comparable-status reconciliation.

Reader	Native L2	Prior L3
GPT-6.2 Sol	30/49	33/50
GPT-6.1 Terra	20/48	24/49
Qwen3.5-9B	0/50	0/50
Grok 4.7 Build	2/32	2/30

Each historical coding cell also had 50 assigned tasks. Native Grok transport-only counts were 18 for L2 and 20 for prior L3. The line-feed-only diagnostic resolved 3/10 and 4/9 eligible Grok outputs, and 0/40 and 0/42 eligible Qwen outputs, for those two arms respectively.

Against prior L3 on shared comparable tasks, L3.1-only resolutions, prior-L3-only resolutions, and ties were 2/2/46 for Sol, 9/7/34 for Terra, 0/0/50 for Qwen, and 0/0/25 for Grok. That final Grok comparison is restricted to 25 commonly scorable tasks.

References

- Islam and colleagues. LEDGER: Auditable Agent Memory Under Change. 2026. Paper
- Snap Research. LoCoMo pinned release and raw dialogue data. Repository
- Wu and colleagues. Cleaned LongMemEval-S release. Dataset
- Wu and colleagues. LongMemEval-V2 Small benchmark and upstream deterministic scorer. Dataset · Scorer repository
- Tongyi-ConvAI. VibeMemBench earlier-trajectory task dataset and official evaluation. Paper · Dataset
- Aletheia. Raw-turn memory benchmark. Published retrieval results · Pinned implementation
- AgentMemory. BM25 SearchIndex. Published retrieval receipts · Pinned implementation
- BAAI, BGE-small-en-v1.5. Model card · Pinned Xenova ONNX export
- Independent data sources. 1 Mnemoverse · 2 Mem0 top-20 study · 3 OmniMemEval · 4 VibeMemBench tables · 5 LongMemEval-V2 Small