PURETENSOR
INTELLIGENCE DISTILLED

A Memory Layer With External Anchors

LongMemEval and LoCoMo on a sovereign retrieval stack

Paper: PT-R-2026-014 v1.0

Author: PureTensor Ltd

Date: 27 August 2026

Status: Published

Abstract

Long-term conversational memory presents a retrieval problem with three coupled axes: answer accuracy, context-token budget, and latency. Published benchmarks often report a single accuracy figure without stating the reader model, the token cost, or the retrieval time, which makes comparison across systems difficult. This paper describes our evaluation of a sovereign retrieval stack against two public benchmarks, LongMemEval-S (500 questions) and LoCoMo (1,540 questions), with every configuration reported on all three axes and the reader model named.

On LongMemEval-S, two zero-cost changes to the answer prompt (chronological ordering of retrieved chunks and an explicit supersession rule for conflicting facts) raised accuracy from 67.4% to 72.8% when the reader and judge were both a 27B open-weight model running entirely on our own hardware. The largest gain appeared in knowledge-update questions (+11.5 points) and single-session-preference questions (+26.7 points). When the reader was switched to gpt-5.4 at medium reasoning and the retrieval window widened to top-25, accuracy reached 90.6% (453 of 500), which exceeds Zep's published 90.2% at a cost of 2.7 times their median context budget. At a comparable token budget (top-10, median 4,988 tokens versus Zep's 4,408) we scored 87.8%, 2.4 points behind. Retrieval latency at either k was approximately 98 ms at the 95th percentile, against Zep's published 162 ms, though our figure excludes any network hop while theirs includes a cloud round trip.

On LoCoMo, raw-chunk retrieval at top-50 scored 82.5% with a Claude Sonnet 5 reader. Adding a distilled fact layer (atomic facts extracted once per session by the sovereign 27B model, then indexed alongside raw chunks) raised accuracy to 86.2% at a median context of 6,142 tokens. Distilled facts alone, without raw chunks, scored 77.6%, a 4.9-point loss that illustrates the fidelity risk of extraction-only architectures. Zep's published LoCoMo score is 94.7% with a gpt-5.4-tier reader; the gap is attributed to full-corpus coverage on a small benchmark, documented gold-answer defects, and a reader mismatch, so this is not a like-for-like comparison.

The results support three claims within the stated limits. First, prompt-level changes that cost nothing at inference can recover double-digit accuracy on question types that depend on temporal order or user preference. Second, a fact-distillation layer helps as a supplement to raw chunks but harms accuracy when used alone. Third, a sovereign configuration with open-weight models can reach the low-70s on LongMemEval-S, which is competitive with the 60–70% range vendors have reported using GPT-4-class readers.

1. Motivation and System Description

Our agent-memory substrate stores long-running conversational context in Markdown topic files, which serve as the source of truth. A Postgres store with vector and full-text indexes acts as a rebuildable cache over these files. Retrieval is hybrid: BM25 lexical matching runs in parallel with dense-vector search using qwen3-embedding-4b embeddings, and the two result lists are fused by reciprocal rank fusion. A cross-encoder (bge-reranker-v2-m3) then reranks the fused list, and the top-k chunks are selected for the answer prompt.

Prior to this work we had no published score on any public long-term-memory benchmark. External comparison was therefore impossible, and internal regression tests could not be calibrated against a known baseline. The motivation for this study was to establish that calibration.

All embedding, reranking, and (in the sovereign configuration) answering and judging run on our own GPUs. Nothing leaves the premises in that configuration. When frontier-tier readers are used, the retrieval path remains local and only the final answer call is sent to an external API.

2. Benchmark 1: LongMemEval-S

LongMemEval-S comprises 500 questions drawn from synthetic multi-session conversations. We ingested the corpus into an isolated benchmark database: 100,548 chunks across 19,206 sessions. Ingest used all four fleet GPUs in parallel with a work-stealing queue, completing in 441 seconds end-to-end (228 chunks per second) against a 45 chunks per second single-endpoint baseline. The speedup factor was approximately 5x. Sustained GPU utilisation reached 100% on all four cards, whereas the baseline sat between 0% and 20% because the client loop was the bottleneck.

Two engineering notes emerged. First, batches of 128 documents at roughly 12 KB each sent approximately 400,000 tokens per request and starved the server queue; one queue reached 716 waiting requests. Dropping the embed batch size to 32 resolved the starvation. Second, the serving engine continues processing requests whose client has already timed out, so retries snowball unless the batch is sized to the server's throughput.

3. Sovereign Configuration Results

The sovereign configuration used our on-premises Qwen3.8-27B (FP8) as both the answer model and the judge. Two runs were conducted on the same database: a baseline and an improved run incorporating two changes. First, the top-10 reranked chunks are now sorted by session date before being assembled into the answer prompt. Rerank order is kept for selection and discarded for presentation, because temporal questions depend on date sequence rather than relevance score. Second, the answer prompt gained three instructions: excerpts are chronological and later excerpts supersede earlier ones when facts conflict; reason from session dates when the question involves time; tailor suggestions to preferences the user expressed in the excerpts.

Both runs completed with zero errors. Results by question type follow.

Question type	Baseline	Improved	Change
Overall	67.4% (337/500)	72.8% (364/500)	+5.4 pts
knowledge-update	78.2% (61/78)	89.7% (70/78)	+11.5 pts
single-session-preference	20.0% (6/30)	46.7% (14/30)	+26.7 pts

Question type	Baseline	Improved	Change
multi-session	63.9% (85/133)	67.7% (90/133)	+3.8 pts
temporal-reasoning	51.9% (69/133)	54.9% (73/133)	+3.0 pts
single-session-assistant	94.6% (53/56)	96.4% (54/56)	+1.8 pts
single-session-user	90.0% (63/70)	90.0% (63/70)	unchanged
abstention subset	100% (30/30)	96.7% (29/30)	−1 question

Knowledge-update questions gained most because chronological order plus the supersede rule lets the model identify the newest statement. Preference accuracy more than doubled from the prompt change alone; the baseline had been returning uninformative abstentions. Temporal reasoning gained modestly. The residual is retrieval-side: some questions need context from many sessions and a top-10 window cannot hold it.

Failure analysis of the improved run's temporal-reasoning misses showed that 52 of 60 misses had the full evidence retrieved. These were reasoning failures, not retrieval failures.

4. Frontier-Tier Configuration and External Comparison

To compare against Zep's published numbers, we ran the same 500 questions through the same retrieval stack (hybrid BM25 plus dense, reciprocal rank fusion, cross-encoder rerank, chronological ordering) but with frontier-tier readers and the official LongMemEval judge protocol. Zep is a commercial memory service whose published figures are the most-cited external reference for this benchmark.

Configuration	Ours	Zep (published)
Reader gpt-4o-2024-08-06, official judge prompts, top-10	75.8%	71.2% (arXiv:2501.13956)
Reader gpt-5.4 at medium reasoning, gpt-5.4 chain-of-thought judge, top-10	87.8% (439/500)	90.2% (Zep research page)
Reader gpt-5.4 at medium reasoning, same judge, top-25	90.6% (453/500)	90.2% (Zep research page)

Per-type results across the three configurations follow.

Question type	gpt-4o top-10	gpt-5.4 top-10	gpt-5.4 top-25
Overall	75.8%	87.8%	90.6%
temporal-reasoning	74.4%	89.5%	93.2% (124/133)
multi-session	60.2%	77.4%	84.2% (112/133)
knowledge-update	93.6%	92.3%	91.0% (71/78)
single-session-user	90.0%	94.3%	94.3% (66/70)
single-session-assistant	92.9%	98.2%	98.2% (55/56)
single-session-preference	40.0%	80.0%	83.3% (25/30)
abstention subset	not recorded	86.7%	90.0% (27/30)

The reasoning reader removed most of the temporal-reasoning failure class: +15.1 points on that type at top-10. The residual gap to Zep's 90.2% at top-10 was concentrated in multi-session questions, consistent with Zep's deeper retrieval budget (approximately 4,408 tokens median context versus our top-10 window). Widening to top-25 gained a further 6.8

points on multi-session and took the overall to 90.6%.

Zero API errors occurred across all 2,000 reader and judge calls. Combined external cost was approximately $12.

5. Token Efficiency and Latency

Context tokens were counted with the o200k_base tokenizer over all 500 questions. Latency was measured from a sequential 100-question pass so that figures are per query, not artefacts of GPU contention.

Metric	Ours, top-10	Ours, top-25	Zep (published)
Accuracy (gpt-5.4 tier)	87.8%	90.6%	90.2%
Median context tokens	4,988	11,699	4,408
Retrieval latency p95	~98 ms	98 ms	162 ms

The fair summary is this: at Zep's token budget (top-10, median 4,988 tokens versus their 4,408) we score 87.8%, 2.4 points behind. Beating their accuracy (90.6% at top-25) costs 2.7 times their context budget. Retrieval latency is lower at either k, with the caveat that ours runs on local hardware with no network hop while Zep's published figure includes their cloud round trip. Any comparison must quote all three axes per configuration, never the best of each.

6. Benchmark 2: LoCoMo

LoCoMo comprises 1,540 questions over 272 conversations, each conversation approximately 26,000 tokens. An ingest-time extraction system like Zep's can cover the whole corpus inside its context budget, while top-k raw-chunk retrieval samples roughly 17% of it. This is a coverage trade, not a retrieval-quality trade.

Raw-chunk results at the gpt-5.4 tier formed a ladder: top-25 scored 74.5%, top-50 scored 79.2%, and top-100 scored 85.5%. All runs were error-free; the top-100 run exhausted the external API credit.

LoCoMo carries documented gold-answer defects (for example, a statement containing "last Saturday" whose gold answer says Sunday) and 96 subjective open-domain questions graded against curator opinion.

7. The Distilled Fact-Twin Layer

Each session was run once through the sovereign 27B model to produce exhaustive, date-anchored atomic facts, packed at fact-group granularity (approximately 1,500 bytes per chunk). Session-sized packing measured worse than raw at a 24,700-token median context. Distilled chunks live on the same artefact under a high chunk index and compete in the same BM25, dense, and rerank pool. We call this "mixed" mode.

Distillation of 272 sessions completed in 281 seconds with zero failures and zero external cost. This step corresponds to what a hosted service meters as ingest credits.

Validation used Claude Sonnet 5 on an EU-region managed endpoint with medium adaptive thinking, after the external credit ran out. Internal deltas are reader-consistent.

Configuration	Overall	multi-hop	temporal	single-hop	Median context tokens
raw, top-50	82.5%	70.6%	80.7%	88.9%	~4,200

Configuration	Overall	multi-hop	temporal	single-hop	Median context tokens
mixed, top-50 (raw + facts)	86.2%	75.5%	87.2%	92.2%	6,142
distilled-only, top-15	77.6%	65.6%	79.1%	82.3%	5,071
Zep published (gpt-5.4 tier)	94.7%	not given	not given	not given	5,760

Distilled facts help as a supplement: +3.7 points at a Zep-comparable token budget. Distilled-only loses 4.9 points against raw. Extraction alone drops fidelity, which is the structural risk in extraction-only memory architectures.

Zero API errors occurred across all 9,240 reader and judge calls in this leg.

The remaining gap to Zep's 94.7% is attributed to full-corpus coverage on a small benchmark, gold defects, and subjective grading. Cross-reader caution applies (Sonnet 5 here versus gpt-5.4 in Zep's figure), so this is not a like-for-like comparison.

8. Production Regression Check

A golden-set retrieval evaluation ran against the production vault on the same day as the benchmark runs.

Metric	25 August 2026	18 August 2026
Recall@5	0.8324	0.8453
Recall@10	0.8508	not recorded
MRR	0.7464	0.7642
nDCG@5	0.7662	0.7833
Queries evaluated	543 of 582	not applicable
Latency p50 / p95	126 ms / 139 ms	not recorded

Recall@5 fell 1.3 points. The drop is attributed to dataset drift: the golden set harvests new pairs continuously, and the corpus grew a week of memories between runs.

9. What These Numbers Will Not Carry

Several limits constrain the claims that can be drawn from this study.

First, all LongMemEval-S and LoCoMo runs are single-run evaluations. We did not repeat runs to measure variance, so the reported accuracy figures carry sampling noise that is not quantified. Small differences (a few points) may be within noise.

Second, the comparison to Zep uses vendor self-reported numbers. We reproduced their protocol on our side only; we did not run their system. Their published figures may reflect different corpus preprocessing, different prompt templates, or different judge behaviour.

Third, the LoCoMo comparison is not like-for-like. Our validation runs used Claude Sonnet 5 while Zep's published figure used a gpt-5.4-tier reader. Reader differences can shift accuracy by several points, so the 8.5-point gap (86.2% versus 94.7%) cannot be attributed solely to retrieval quality.

Fourth, latency comparisons are confounded by network topology. Our retrieval latency excludes any network hop because the system runs on local hardware. Zep's published 162 ms

includes their cloud round trip. The figures are not directly comparable.

Fifth, the production regression check shows a 1.3-point drop in Recall@5 over one week. This is attributed to dataset drift, but we cannot rule out other causes because the comparison is observational, not controlled.

Sixth, LoCoMo's gold-answer defects and subjective questions introduce grading noise that is not under our control. Some fraction of misses may be correct answers penalised by flawed gold labels.

The claims that survive these limits are: (a) chronological ordering and the supersede prompt rule improve accuracy on knowledge-update and preference questions by double digits in the sovereign configuration; (b) distilled facts help as a supplement but harm accuracy when used alone; (c) at a comparable token budget to Zep's published top-10 configuration, we score 2.4 points behind on LongMemEval-S; (d) widening to top-25 exceeds their accuracy at 2.7 times the token cost.

10. Next Steps

Two changes have already shipped in the memory engine. Chronological ordering of retrieved chunks is now the default before prompt assembly, and the supersede and preference-tailoring instructions are part of the answer prompt.

The next experiment is a repeat of the LoCoMo mixed-mode run with a gpt-5.4-tier reader, so that the distilled-fact result can be placed against Zep's published figure without the reader mismatch that currently prevents a like-for-like comparison.

PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. Measurements for this paper were taken on 25 August 2026; raw artefacts are retained. This paper is PT-R-2026-014.