Local Memory-System Benchmark

Companion Memory Server vs. Mem0, Hindsight, OpenViking, Holographic, and Naive RAG

Companion Intelligence — Companion Memory Server

June 2026

© 2026 Lifescope Inc. All rights reserved.

Questions and data requests: research@ci.computer

Abstract

A fully local, head-to-head evaluation of six agent-memory systems on a single synthetic persona, with no cloud calls.[1] On a 100-question flat/relational corpus the top three are a statistical tie — **naive 79, mem0 78, Companion Memory Server 75** (answer accuracy/100) — followed by **Holographic 69, OpenViking 47, and Hindsight 25**. On *structured* tasks (relationship enumeration, temporal supersession), the workloads a knowledge graph is built for, **Companion Memory Server leads**: best temporal, best enumeration, best retrieval completeness, at top-tier query latency. Plain vector RAG is brutally cost-effective for one-shot fact lookup; the graph's edge appears on relationship traversal and "what is true now."

1 Method

Corpus. One synthetic persona ("Max Smith"): **90 hand-authored canonical facts** (people, places, employers, relationships, multi-hop chains, facts that change over time), buried under **400 distractor messages** (the same topics about *other* people, plus low-signal filler) so a fixed top-k retriever cannot trivially win. Total **490 messages; top-10 ∼2% coverage**.

Uniform contract. Each system runs in its own container exposing the same HTTP surface (`reset`, `ingest`, `flush`, `retrieve`, `answer`); the orchestrator speaks only this contract.

Two-phase, fair by construction.

- *Phase A — ingest/extraction:* every system ingests the identical corpus using the same extractor LLM (`gemma-4-31b-it-abliterated`) and the same embeddings (`nomic-embed-text`).[2] This isolates the memory *architecture*, not the model.
- *Phase B — answer + judge:* a single, constant retrieval-augmented answerer and LLM judge (`ornstein-3.6-27b`) is used for *all* systems, so we compare memory layers, not answer-writers.

[1] Cloud-only providers (Honcho, RetainDB, Supermemory, Memori) were excluded by the local-only constraint. Mem0 was run self-hosted (open-source library + Chroma) against local models, not its hosted SaaS.

[2] Hindsight uses its built-in `bge-small` embedder, a flagged minor confound. Holographic uses *no* LLM and *no* embeddings (SQLite FTS5 + HRR).

Two scored dimensions. *Answer*: did the constant answerer, given the system's retrieved context, produce a correct final answer? *Storage*: was the gold fact actually present in the *raw retrieved context*?[3]

2 Question taxonomy

Three suites span nine classes. Every gold answer is grounded in exactly one canonical fact, so a miss is a retrieval/answer failure, not a missing fact.

Suite A — Broad recall (100 questions)

`flat-recall` **(34)** A single atomic fact stated once. *"What instrument does Max play?"* → violin. Probes whether the system surfaces one stated fact from under 400 distractors. LLM-judged.

`relational` **(22)** One relationship hop. *"Where do Max's parents live?"* → Portland.

`multi-hop` **(22)** Two or more hops. *"What does Max's brother-in-law do for work?"* → cardiologist (Max → sister Anita → husband Raj → occupation).

`temporal-update` **(22)** Facts that changed; asks for the current value. *"What city does Max currently live in?"* → Denver, not the prior Portland. Probes supersession.

Suite B — Structured / agentic (37 questions)

`temporal-update` **(20)** As above, both directions (current *and* prior value).

`enumeration` **(10)** "List all X" set queries. *"List all of Max's family members."* Scored by **deterministic set recall** — the fraction of gold members surfaced — with no LLM judge.[4] Probes returning the *complete* typed set, the graph's structural strength.

`aggregation` **(7)** Counts and single-target filters. *"How many cities has Max lived in?"* → 2. LLM-judged.

Suite C — Absence / anti-hallucination (23 questions)

`absence` **(16)** Baited negatives; the answer is *no*, and each is baited with a tempting distractor about someone else. *"Does Max work at Stripe?"* → No (that is his friend Lena's employer). Probes hallucination resistance. Scored by an existence-aware judge.[5]

`presence-control` **(7)** Genuine "yes" questions, so a system that blindly answers "no" cannot game the absence suite.

[3] Storage isolates "did the memory store + surface the fact" from "did the answerer phrase it well." A system can have high storage but lower answer, or vice versa.

[4] Set recall is reproducible and unspinnable: it counts how many of the canonical members appear, in the answer and (for storage) in the retrieved context.

[5] Correct only if the yes/no matches truth; a confident "yes" to a non-existent thing is a hallucination (wrong); "I don't know" is treated as not-a-hallucination.

3 Results

Table 1: Suite A — 100 questions (answer / storage, then per-category answer).

System	Answer	Storage	flat (34)	relational (22)	multi-hop (22)	temporal (22)
naive	**79**	78	28	**22**	15	14
mem0	78	**79**	24	21	**19**	14
Companion Memory Server	75	78	26	19	12	**18**
holographic	69	72	25	20	13	11
openviking	47	56	18	6	15	8
hindsight	25	21	7	9	4	5

Table 2: Suite B — Structured (answer / storage; enumeration scored by set recall).

System	Answer	Storage	temporal (20)	enumeration (10)	aggregation (7)
mem0	**25.2**	26.2	13.0	6.2	**6.0**
naive	24.9	24.2	14.0	4.9	**6.0**
Companion Memory Server	24.1	**27.0**	**15.0**	**7.1**	2.0
openviking	12.8	14.4	7.0	2.8	3.0
hindsight	9.9	10.2	5.0	2.9	2.0
holographic	9.5	11.5	7.0	0.5	2.0

Table 3: Suite C — Absence / anti-hallucination (answer / absence / presence-control).

System	Answer	absence (16)	presence-control (7)
naive	20	15	5
mem0	18	14	4
Companion Memory Server	18	11	**7**
hindsight	18	14	4
holographic	20	14	**6**
openviking	15	12	3

Table 4: Cost — memory-layer work only. The answer phase is the shared model (excluded as a per-system cost).

System	Ingest (490 msgs)	Recall latency/query	Concurrency-safe
holographic	**5 s** (no LLM/embeddings)	∼5 ms	yes
naive	16 s	∼60 ms	yes
mem0	∼15 m	∼120 ms	yes
Companion Memory Server	∼16 m	∼80 ms	yes
openviking	∼50 m (VLM extraction)	∼30 ms	yes
hindsight	∼35 m[6] / ∼2.3 h seq	∼1160 ms	no[7]

Recall latency is the median of 3 warm `/retrieve` *calls; naive's ∼1 s cold first call is excluded.*

3.1 Charts

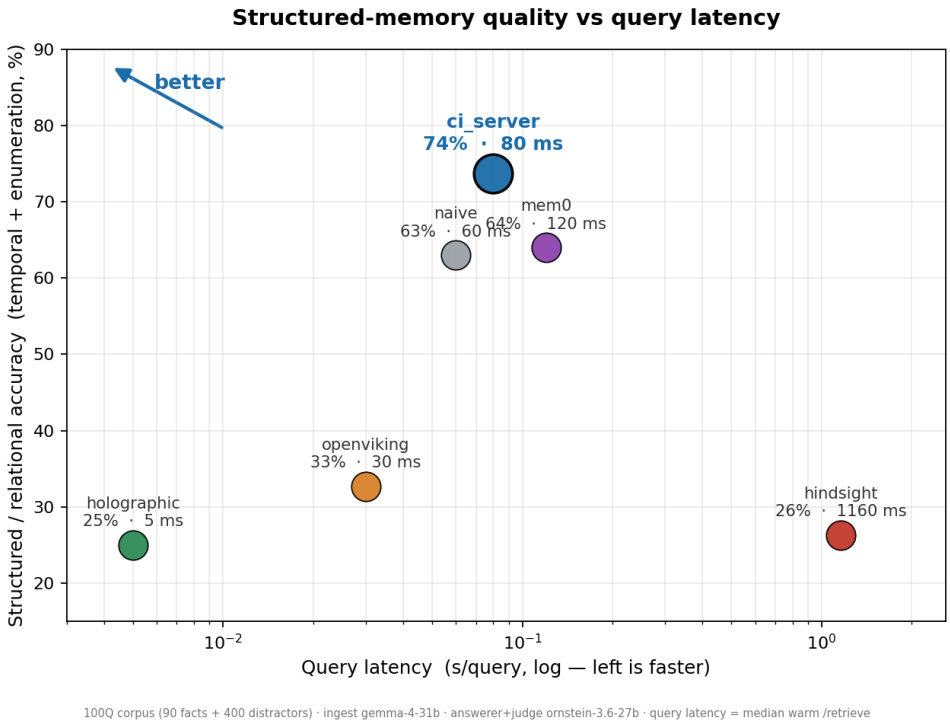

Figure 1: Structured/relational accuracy (temporal + enumeration) vs. query latency. Companion Memory Server occupies the top-left (highest structured quality at top-tier latency). Holographic and OpenViking are fast but low on structured quality; Hindsight is slow and low.

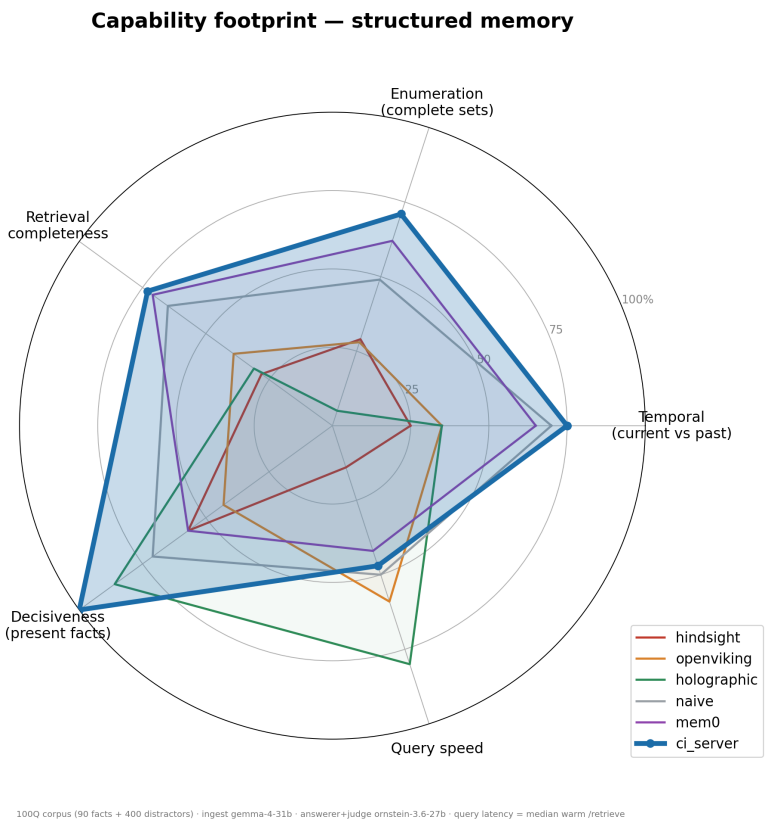

Figure 2: Capability footprint across five axes. Companion Memory Server has the largest enclosed area, leading temporal, enumeration, retrieval completeness, and decisiveness on present facts.

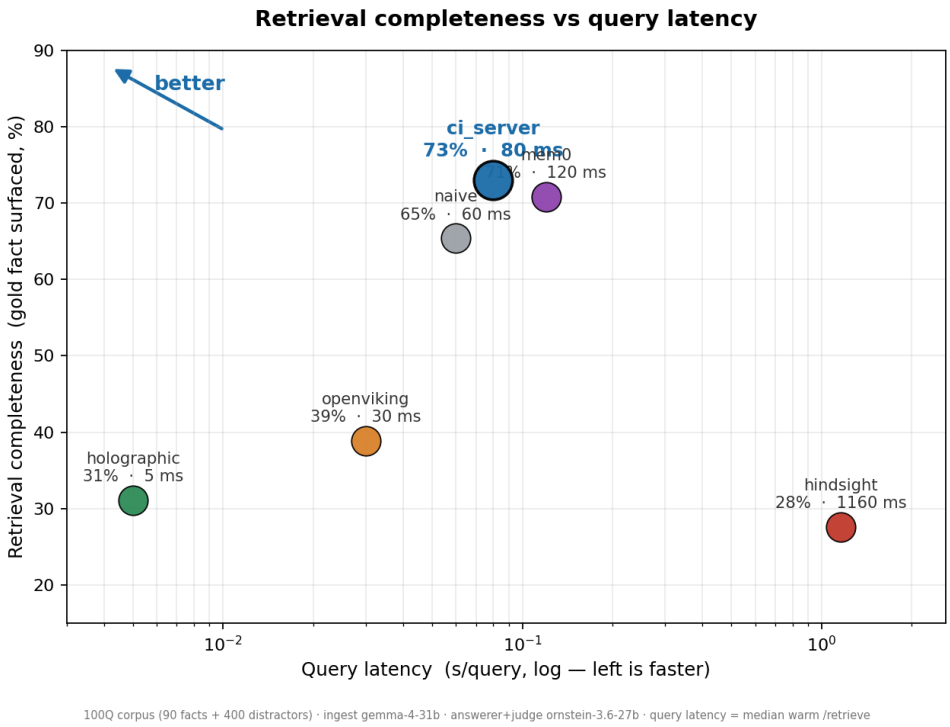

Figure 3: Retrieval completeness (fraction of gold facts actually surfaced) vs. query latency. Companion Memory Server is highest on completeness.

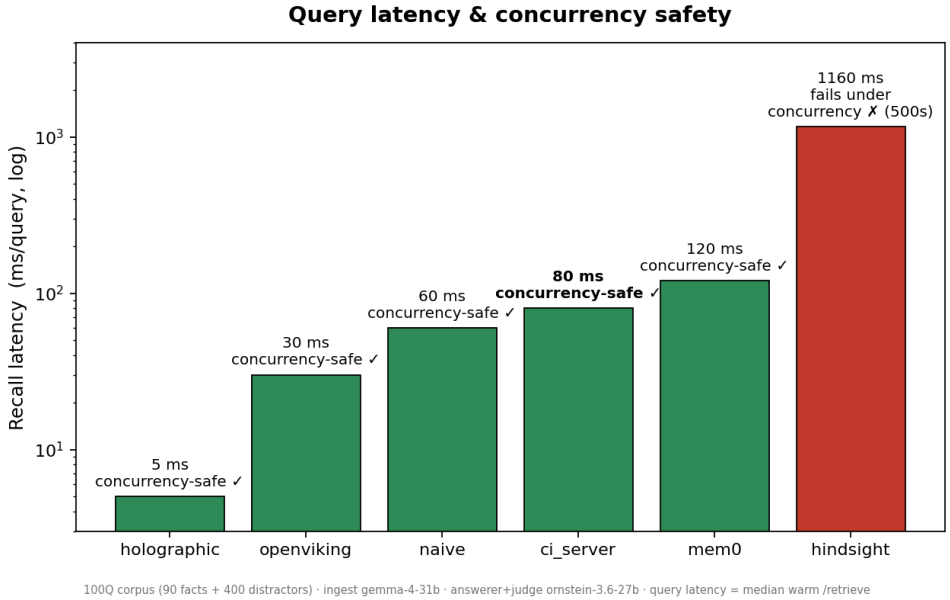

Figure 4: Recall latency and concurrency safety. Five of six systems are concurrency-safe; Hindsight fails under concurrent recall.

4 Conclusions

- **Flat fact lookup is a solved problem.** Two lightweight systems — naive (vector) and Holographic (keyword/HRR[8]) both reach ∼70–79 at trivial cost. A graph is overkill here.
- **Companion Memory Server's edge is structured / temporal / relational memory:** best temporal, best enumeration, best retrieval completeness, most decisive on present facts, at ∼80 ms recall and concurrency-safe.
- **Trade-offs:** aggregation (counting) and absence-precision are weak spots (the flip side of surfacing rich context); multi-hop trails Mem0's consolidation; ingest is mid-pack.
- **The other graph/context-DB systems land below Companion Memory Server on structured tasks.** OpenViking[9] is fastest at recall but slowest to ingest; Hindsight is slowest and weakest overall.
- **Holographic is the efficiency dark horse:** fastest ingest (5 s) and recall (∼5 ms), 69/100 on broad recall — but it *collapses* on enumeration (0.5/10): a keyword store cannot assemble a set, exactly where Companion Memory Server's structure wins.

Reproduce

All artifacts live in `scripts/benchmark/v2/`: `orchestrator.py` (harness), `docker-compose.yml` (one container per system), personas in `../personas/`, per-suite reports (`report_v2_100q.md`, `report_v2_struct.md`, `report_holo_*.md`), and charts in `charts/` (regenerate with `make_charts.py`).

[8] Holographic is keyword-only (SQLite FTS5 + HRR), with no semantic embedder. Its FTS5 layer ANDs query terms, so a raw natural-language question matches nothing; the adapter sends an *OR-of-keywords* query (stopwords dropped), the same treatment naive and Companion Memory Server's boosted search already apply to their keyword terms. Without it Holographic scores ∼0 — an artifact of query format, not the system. Vendored from `NousResearch/hermes-agent` plugins/memory/holographic; the Hermes-runtime plugin glue was replaced by a thin adapter.

[9] OpenViking (Volcengine/ByteDance) is a self-hostable "context database"; we ran its server on the same local models (gemma-4-31b VLM + nomic embeddings). Its text-only-VLM path works on our image-free corpus. License: AGPL-3.0 — a consideration for shipping, not for benchmarking.