

Remember Everything, Inject Almost Nothing: Results for a Long-Term Memory Layer on LoCoMo, LongMemEval and BEAM

Madhu

Independent · Kansas City · madhudream.dev

ABSTRACT

We report results for hanumemAI, a long-term memory layer for language-model assistants, on three public benchmarks, each run on its full question set. On LoCoMo (1,540 questions) it scores 96.1, the highest published figure we are aware of, while placing 1,533 context tokens per question in front of the answering model; the published systems closest to it read 6,700 to 45,800. On BEAM at one million tokens (700 questions) it scores a mean rubric of 68.5 at 2,716 tokens, against a published 64.1 at 6,700 for Mem0. On LongMemEval-S (500 questions) it scores 94.0 at 2,428 tokens with a small open-weight answering model, where every other published result of 93 or more uses a frontier model; it is not first on that benchmark, nor on BEAM. We report each score with its answering model, its judge and its context cost, grade the same answers under two judges, and show that the choice of judge alone moves a score by two to five points. The results come from 162 logged runs, of which 23 changes were kept. The system is proprietary: this paper reports what it measures and not how it is built.

1 Introduction

A language model keeps nothing between one call and the next. What it knows about a user is what is in its prompt. An assistant that is to know someone across months therefore needs a system outside the model that decides, for every question, which few lines of everything it was told are placed in front of the model.

Such a system can always buy accuracy with context. Published results for memory systems range from about 1,600 to about 46,000 context tokens per question (Section 3), and the scores are rarely quoted with that number beside them. We take the position that the result of a memory layer is a pair, accuracy and context tokens per question, and that a score quoted without its answering model, its judge and its context cost cannot be compared with anything.

This is a results paper. It states what was measured, on which questions, with which models, and where the system is behind. It does not describe the design.

2 Benchmarks and protocol

2.1 Benchmarks

LoCoMo [1] is ten very long conversations of about 300 turns each, with 1,540 questions in the four categories that are normally scored: single-hop, multi-hop, temporal and open-domain. **LongMemEval-S** [2] is 500 questions of six types and abstention, each hidden in about 115,000 tokens of chat sessions. **BEAM** [3] is conversations at four sizes from about 128,000 to ten million tokens, with questions over ten abilities graded against a rubric; we report the one-million split in full (35 conversations, 700 questions) and the 100K split (20 conversations, 400 questions).

2.2 What every score is reported with

The answering model (the reader); the judging model and its prompt; the number of questions; and the mean number of context tokens the reader was given per question. For BEAM we report the mean rubric score, which is what other systems publish, and the pass rate (rubric of 0.5 or more) beside it. The LoCoMo judge prompt is the one in Mem0's open harness, unchanged; the LongMemEval judge prompts are the paper's own.

2.3 Development

Each change to the system was one experiment: a single change, measured on all three benchmarks, and kept only if the score held on each. The log has 162 lines: 23 changes kept, 40 discarded, and 99 measurements that were not candidates.

3 Results

Figures for other systems in this section are as published by their vendors, collected on 28 September 2026. None has been reproduced by us or, to our knowledge, by a third party.

3.1 LoCoMo

System	Reader	Judge	Acc.	Tokens / Q
hanumemAI	qwen3.8-flash, thinking	gpt-4o-mini	96.1	1,533
hanumemAI, same answers	qwen3.8-flash, thinking	qwen3.7-flash	91.6	1,533
Agent Zero	Claude Sonnet 5	not named	93.6	45,800
EverMemOS	gpt-4.1-mini	gpt-4o-mini + 2	93.05	~6,700
Mem0 platform	gpt-5	gpt-5	92.5	7,000
Hindsight	gemini-3.1-pro	gemini-2.5-flash-lite	92.0	36,200

System	Reader	Judge	Acc.	Tokens / Q
MemMachine	gpt-4.1-mini	gpt-4o-mini	91.7	–
Mem0 paper [4]	gpt-4o-mini	–	66.9	~7,000

Table 1: LoCoMo, all 1,540 questions. By category, the first row scores 97.6 single-hop, 95.7 multi-hop, 95.3 temporal and 86.5 open-domain.

3.2 LongMemEval-S

System	Reader	Acc.	Tokens / Q
agentmemory V4	Claude Opus 4.6	96.2	1,500–7,500
Agent Zero	gpt-5.5	95.6	12,200
Chronos High	Claude Opus 4.6	95.6	–
Hindsight	gemini-3.1-pro	94.6	43,600
Mem0 platform	gpt-5	94.4	6,800
hanumemAI	qwen3.8-flash, thinking	94.0	2,428
Mastra OM	gpt-5-mini	93.6	~30,000
MemMachine	gpt-5-mini	93.0	~19,600
SodaMem	deepseek-v4-flash	92.8	18,300
HydraDB	Gemini 3 Pro	90.8	–
Honcho	Claude Haiku 4.5	90.4	–
Zep	gpt-4o	71.2	~1,600

Table 2: LongMemEval-S, all 500 questions. Our row is judged by qwen3.7-flash with the paper’s prompts; the same answers score 91.8 under gpt-4o-mini. By type: single-session user 96.9, single-session assistant 100, preference 96.7, knowledge update 100, temporal 96.9, multi-session 84.3, abstention 86.7.

3.3 BEAM

System	Reader	100K	1M	Tokens / Q
Hindsight	gemini-3.1-pro	73.4	73.9	17,700–27,300
hanumemAI	qwen3.8-flash, thinking	70.5	68.5	2,716–3,053
Mem0 platform	gpt-5	–	64.1	6,700
Honcho	Claude Haiku 4.5	63.0	63.1	–
Benchmark paper baselines [3]		32–36	31–34	

Table 3: BEAM, mean rubric score. Our pass rates are 76.5 at 100K and 73.4 at 1M. At 1M by ability: preference 98.6, instruction 85.7, information extraction 81.4, multi-session 75.7, contradiction 74.3, abstention 71.4, knowledge update 71.4, temporal 64.3, event ordering 58.6, summarization 52.9.

3.4 Accuracy against context

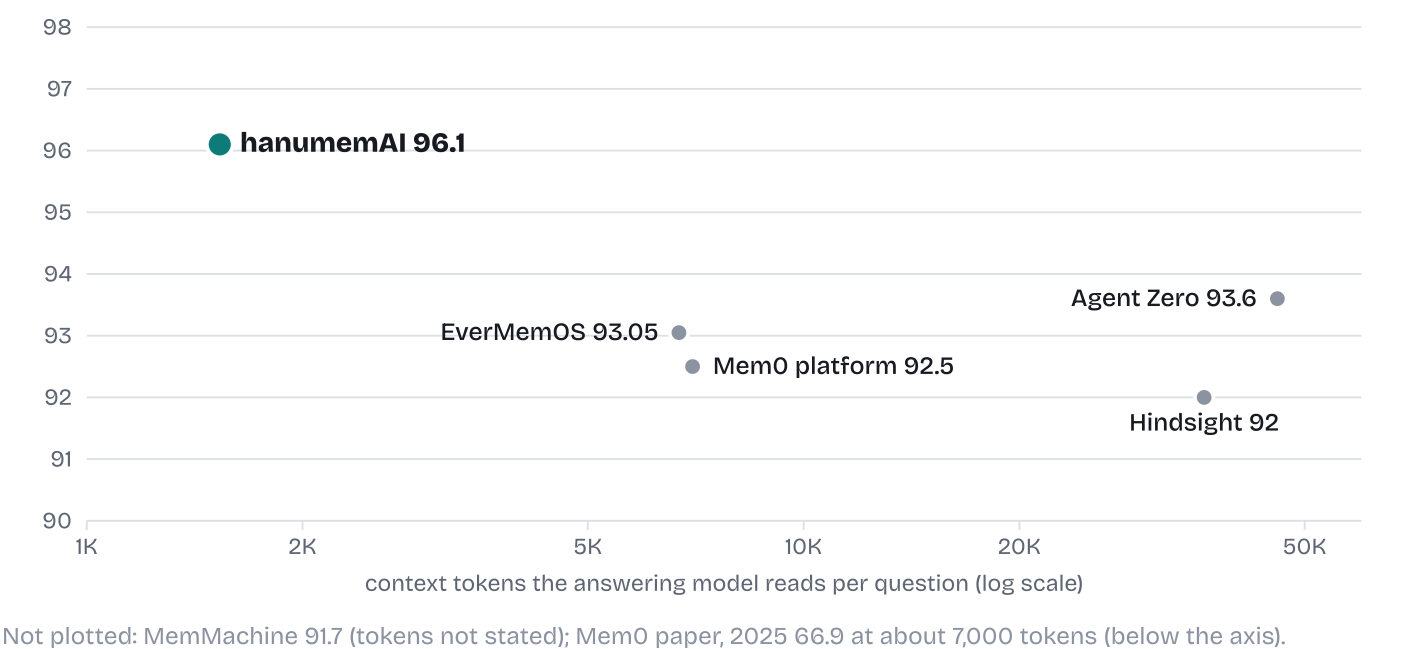


Figure 1: LoCoMo accuracy against context tokens per question, log scale, for every system that states both. Up and to the left is better.

Across the three benchmarks the system reads between 1,533 and 3,053 context tokens per question. The systems that score near or above it read between 6,700 and 45,800. On BEAM-1M it is ahead of Mem0 by 4.4 points at 40 percent of the tokens, and behind Hindsight by 5.4 points at 11 percent of the tokens.

4 What moves a score besides memory

4.1 The judge

We graded identical answers under two judges. On LoCoMo, one set of 1,540 answers scored 88.5 under qwen3.7-flash and 94.0 under gpt-4o-mini; the final run scores 91.6 and 96.1. On LongMemEval-S the order reverses: 94.0 under qwen3.7-flash and 91.8 under gpt-4o-mini. Neither judge is wrong. They draw the line for a paraphrase in different places, and which is stricter depends on the benchmark. We therefore publish both.

4.2 The reader

On LongMemEval-S, with the stored memories unchanged, the score was 81.0 with qwen3.7-flash as reader and 88.6 with qwen3.8-flash. A larger reader is not always better: with gpt-5 as reader on the same memories and the same judge, a 100-question LongMemEval sample scored 94.0 against 95.0 for the small thinking reader, and 100 BEAM-100K questions scored a rubric mean of 67.0 against 69.6. The samples are small, and the answer prompt was tuned with the small reader, so this measures gpt-5 inside this stack and not gpt-5 at its best.

4.3 Samples and noise

A 100-question sample of LongMemEval scored 91 where the full 500 scored 81.0 with the same code. One configuration run twice on 385 LoCoMo questions scored 93.51 and 92.47. All headline figures here are on full question sets.

5 Limitations

The system is not first on LongMemEval-S (96.2 is the highest published figure) or on BEAM (Hindsight, 73.9 at 1M). No two vendors share a judge, so the tables are context and not a ranking; none of the figures, ours included, has been independently reproduced. There is no held-out test set: the experiments were tuned on the question sets they are scored on. Failures were read on 179 of the 500 LongMemEval questions only, and the 321 that were never read improved by a similar amount (78.5 to 84.4), which is evidence against fitting to the questions but is not a test set.

6 Open problems

Multi-session questions on LongMemEval-S, which require counting across conversations (84). At one million tokens on BEAM: summarization (53), event ordering (59) and temporal reasoning (64). At 100K: abstention and knowledge update (62.5 each).

7 Availability

The design of the system, its prompts and parameters, and the description of each experiment are not in this paper. The subject is taught from zero in *AI Agent Memory: Novice to Ninja* [5]. The library, the run log, the full book and a full technical paper are available on request to the author.

References

1. Evaluating Very Long-Term Conversational Memory of LLM Agents. 2024.

2. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. 2024.
3. Beyond a Million Tokens: Benchmarking and Enhancing Long-Term Memory in LLMs. ICLR 2026.
4. Mem0. 2025. The paper reporting 66.9 on LoCoMo.
5. AI Agent Memory: Novice to Ninja. interactive.carebun.com/books/memory.

Cite

```
@misc{hanumem2026,  
  title = {Remember Everything, Inject Almost Nothing: Results for a Long-Term Memory Layer  
          on LoCoMo, LongMemEval and BEAM},  
  author = {Madhu},  
  year = {2026},  
  url = {https://interactive.carebun.com/papers/hanumem/}  
}
```