THE MEMORY BENCHMARK NOBODY BEATS

Everyone passes the easy tests. *The hard one breaks them.*

LoCoMo and LongMemEval are **saturated** — vendors self-report into the 90s, and on paper a few even beat us. That's the point: high scores there are table stakes. **MemoryAgentBench** (ICLR 2026) was built to be hard to game — and it's where every method collapses. The paper's best system tops out at **49.6**. We score **62.19**.

SATURATED · TABLE STAKES

LOCOMO

84.86

Ours. Vendors self-report up to **92.5%** — numbers they openly dispute.

LONGMEMEVAL

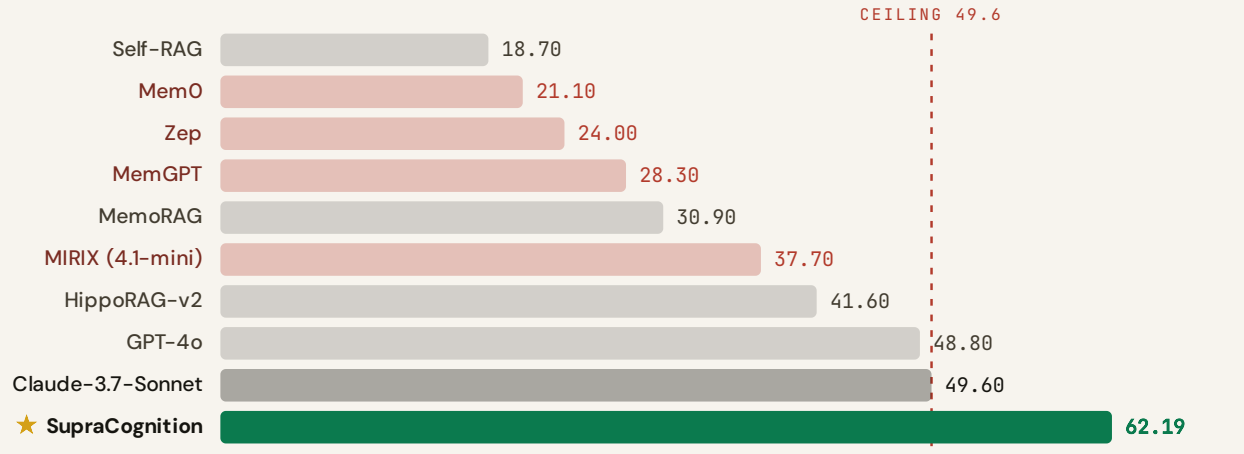
86.0

Ours. Self-reports reach **94.4%**; several systems top us here.

When everyone scores in the 80s and 90s, the benchmark has stopped measuring anything. *High scores here are the price of entry — not a moat.*

THE HARD ONE · MEMORYAGENTBENCH (ICLR 2026)

Where the field collapses, *we cracked the ceiling.*



Overall scores, MemoryAgentBench (Hu et al., ICLR 2026). No method in the paper masters all four memory competencies; on multi-hop conflict resolution, **every system tested collapses to single digits** (≤ 7%). The best result is **49.6**. **SupraCognition scores 62.19** — above the ceiling, and roughly **3×** the dedicated memory systems.

THE COLLAPSE · SAME SYSTEMS, ONE NEUTRAL HARD TEST

SYSTEM	LOCOMO*	LONGMEMEVAL*	MEMORYAGENTBENCH ↓
Mem0	92.5%	94.4%	21.1
Zep	~80%	71%	24.0
SupraCognition	84.86%	86%	62.19

***Self-reported.** Mem0 and Zep publish **80–94%** on the saturated benchmarks — Mem0's numbers sit *above* ours. On the **neutral** hard test they collapse to **21.1** and **24.0**. We score **62.19** — and hold up across all three.

THE BOTTOM LINE

Others post high scores on the easy tests, then collapse on the hardest. SupraCognition scores high on every benchmark tested — and the highest on the one built to break memory systems.

Sources. MemoryAgentBench — Hu et al., "Evaluating Memory in LLM Agents via Incremental Multi-Turn Interactions," ICLR 2026 (arXiv:2507.05257); overall scores and the ≤7% multi-hop conflict-resolution finding from the paper's Table 1 / dataset card. LongMemEval — Wu et al., ICLR 2025 (arXiv:2410.10813). LoCoMo — Maharana et al., 2024. Self-reported figures: Mem0 92.5% LoCoMo / 94.4% LongMemEval (mem0.ai/research, Apr 2026 algorithm); Zep ~80% LoCoMo (latest — originally claimed 84%, corrected to 75.14% in a public dispute with Mem0) / 71.2% LongMemEval (GPT-4o, blog.getzep.com). LoCoMo figures are disputed between vendors. MemoryAgentBench figures from the paper's Table 1; backbones vary across systems; SupraCognition's 62.19 exceeds the highest overall result reported in the paper.