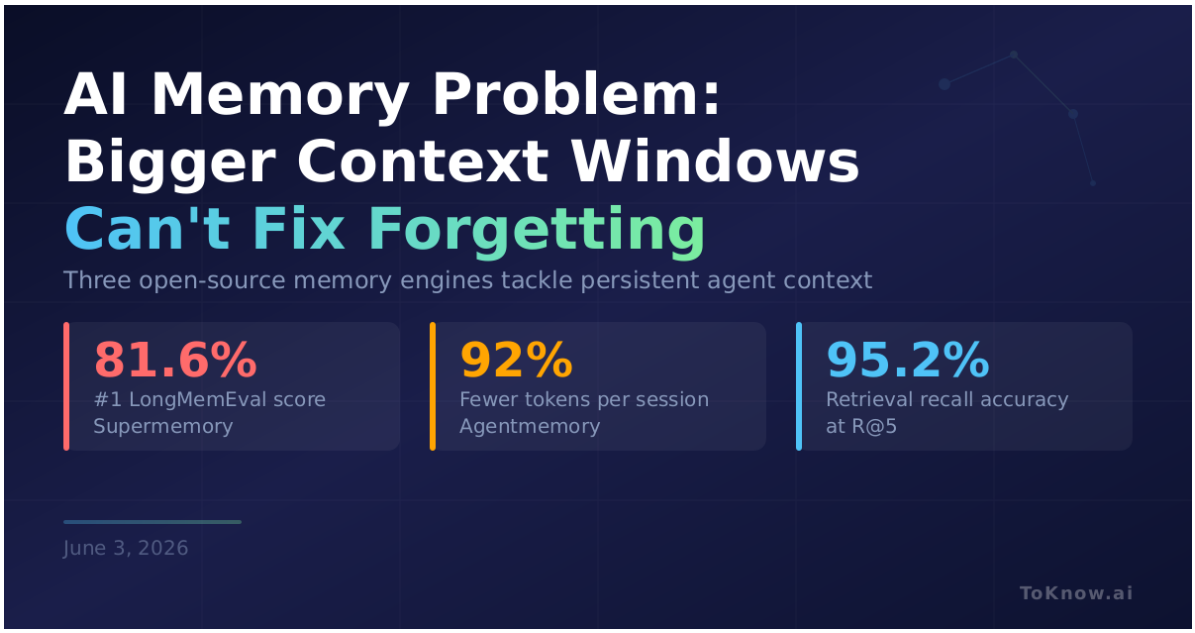
AI Memory Problem: Why Bigger Context Windows Don't Stop Agents From Forgetting

Kabui, Charles

2026-06-03

Read at ToKnow.ai



Even with million-token context windows in Claude and Gemini, AI agents forget everything between sessions and miss information buried in long inputs. A Stanford study confirmed that retrieval degrades sharply when relevant information sits in the middle of long contexts, and bigger windows alone don't fix this. Three open-source projects attack the problem differently. EverMemOS converts conversations into structured memory cells, consolidates them

into thematic scenes, and achieves state-of-the-art on the LoCoMo and LongMemEval benchmarks. Supermemory ranks #1 on all three major memory benchmarks (LongMemEval at 81.6%, LoCoMo, and ConvoMem) with automatic fact extraction and contradiction resolution. Agentmemory targets coding agents: 12 auto-capture hooks record tool usage, compress sessions into a 4-tier memory hierarchy (working, episodic, semantic, procedural), and report 92% fewer tokens with 95.2% retrieval recall.

The cost adds up. At Claude Opus 4.8's $5/$25 per million token pricing, a developer running 50 sessions a day can spend $50-100 just re-reading files the agent already processed. Agentmemory cuts session context to roughly 1,900 tokens from 22,000+. Supermemory's user profiles return in about 50ms. LongTraceRL from Tsinghua attacks from the model side, using reinforcement learning with rubric rewards to train 4B-30B parameter models to navigate long contexts more reliably across five benchmarks.

The context problem is splitting into two engineering disciplines. Longer windows and KV cache compression solve input capacity. Persistent memory systems solve continuity. The memory layer is becoming its own infrastructure stack.

Sources:

- Lost in the Middle: How Language Models Use Long Contexts (Stanford)
- EverMemOS: Self-Organizing Memory Operating System (arXiv)
- Supermemory: Memory Engine for AI
- Agentmemory: Persistent Memory for AI Coding Agents
- LongTraceRL: Long-Context Reasoning from Search Trajectories (Tsinghua)
