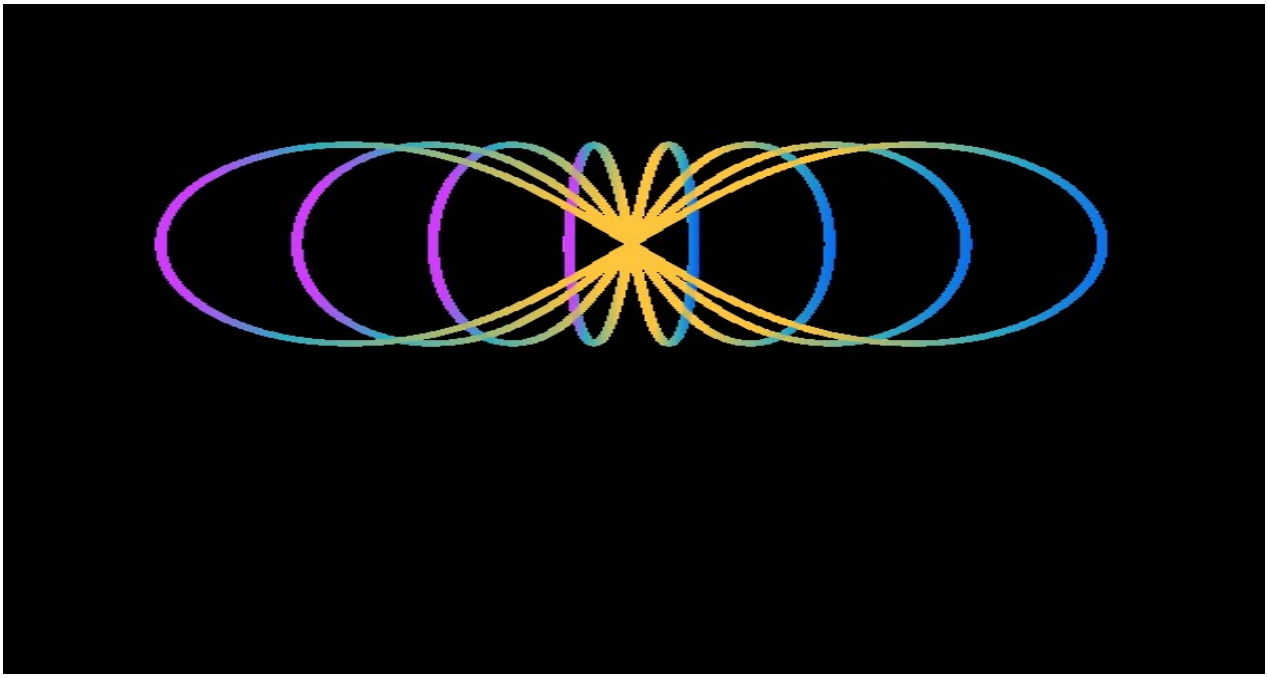


Evermind AI Launches EverMemOS to Transform Artificial Intelligence Through Foundational Memory Infrastructure

December 31, 2025

SAN MATEO, CA - December 31, 2025 -

Modern artificial intelligence systems operate with a fundamental paradox: they demonstrate remarkable reasoning capabilities while simultaneously suffering from systematic amnesia. Large language models can process complex queries and generate sophisticated responses, yet they lack the ability to maintain coherent memory across interactions. Their contextual awareness resets with each new conversation, and their knowledge remains frozen at the point of training. Evermind AI, a team of engineers and researchers focused on AI memory infrastructure, has developed EverMemOS?the Memory OS for Agentic AI?as long-term memory infrastructure for AI that remembers, adapts, and evolves.

Without persistent memory, AI systems cannot achieve genuine personalization because they lack mechanisms to build and retain understanding of individual users over time. They struggle to maintain behavioral consistency, as each interaction exists in isolation from previous exchanges. They cannot self-improve through accumulated experience, as there is no substrate for storing and applying learned patterns. These constraints prevent AI from transitioning from static tools into continuous, adaptive partners capable of long-term collaboration.

"We are not merely adding another layer to the AI stack; we are architecting the foundational memory operating system that will serve as the core data infrastructure for the next era of intelligent agents," said Jason Deng, CoFounder of Evermind AI. "Our mission is to transform AI from brilliant amnesiacs into continuous, personalized partners that grow from every interaction."

EverMemOS addresses these limitations through what Evermind calls a "Layered Memory Extraction" architecture. The system implements a four-layer structure inspired by human memory systems, with each layer corresponding to specific cognitive functions observed in neuroscience. An API and Model Context Protocol interface layer connects with external enterprise systems. An index layer manages embeddings, key-value pairs, and knowledge graph structures, operating comparably to hippocampal memory indexing. The memory layer provides long-term storage and retrieval capabilities similar to cortical memory networks, while the agentic layer handles task understanding, planning, and execution in ways that mirror prefrontal cortex functions.

Rather than treating memory as a simple database for information retrieval, the system transforms memory into an active component that directly influences reasoning processes and output generation. The framework converts raw conversational data into structured semantic units called MemUnits, which are then organized into adaptive memory graphs. This hierarchical extraction and dynamic organization overcomes limitations inherent in similarity-based retrieval methods, providing a stable foundation for maintaining contextual understanding across extended timeframes.

The modular design of EverMemOS allows memory strategies to adapt to different use cases. Enterprise applications demanding precise task execution can implement memory behaviors optimized for accuracy and compliance, while companion AI systems prioritizing relationship continuity can employ strategies that emphasize emotional context and personal history. This flexibility enables the infrastructure to support diverse real-world applications while maintaining optimal memory performance for each specific scenario.

Performance benchmarks demonstrate the practical effectiveness of this approach. On the LoCoMo dataset, which evaluates long-term contextual memory and reasoning capabilities, EverMemOS achieved 92.3 percent accuracy with fully open-sourced and reproducible results. On the LongMemEval-S benchmark, the system reached 82 percent accuracy, substantially outperforming alternative memory solutions tested under identical conditions.

Beyond the core memory operating system, Evermind AI has developed complementary technologies that extend memory capabilities. The company's EverMemModel is approaching the 100 million token context barrier, delivering state-of-the-art performance on tasks such as NQ320k. The proprietary EverMemReRank module more than doubled question-answering performance on the 2Wiki benchmark compared to baseline

models, demonstrating significant practical impact on multi-hop reasoning tasks.

To support broader industry progress, Evermind has created a unified evaluation framework designed to fairly assess modern AI memory systems. The framework tests leading solutions including Mem0, MemOS, Zep, and MemU using consistent datasets, APIs, and metrics that reflect real production performance.

Evermind AI has released EverMemOS as open-source software, enabling developers and organizations to integrate long-term memory capabilities into their own applications. The infrastructure can be deployed through Docker containers and configured to work with various large language model APIs and embedding services. By making this foundational infrastructure publicly available, the company enables experimentation and innovation across the broader AI development community, positioning memory infrastructure as a foundational layer comparable to operating systems in traditional computing.

###

For more information about Evermind AI, contact the company here:Evermind AISophiaevermind@shanda.comSAN MATEO

Evermind AI

EverMind AI is the creator of EverMemOS, the Memory OS for Agentic AI. We provide long-term memory infrastructure for AI that remembers, adapts, and evolves, serving as the foundational core for the next generation of intelligent systems.

Website: https://evermind.ai/

Email: evermind@shanda.com

