What is AI agent memory?

AI agent memory refers to an artificial intelligence (AI) system's ability to store and recall past experiences to improve decision-making, perception and overall performance.

Unlike traditional AI models that process each task independently, AI agents with memory can retain context, recognize patterns over time and adapt based on past interactions. This capability is essential for goal-oriented AI applications, where feedback loops, knowledge bases and adaptive learning are required.

Memory is a system that remembers something about previous interactions. AI agents do not necessarily need memory systems. Simple reflex agents, for example, perceive real-time information about their environment and act on it or pass that information along.

A basic thermostat does not need to remember what the temperature was yesterday. But a more advanced "smart" thermostat with memory can go beyond simple on or off temperature regulation by learning patterns, adapting to user behavior and optimizing energy efficiency. Instead of reacting only to the current temperature, it can store and analyze past data to make more intelligent decisions.

Large language models (LLMs) cannot, by themselves, remember things. The memory component must be added. However, one of the biggest challenges in AI memory design is optimizing retrieval efficiency, as storing excessive data can lead to slower response times.

Optimized memory management helps ensure that AI systems store only the most relevant information while maintaining low-latency processing for real-time applications.

Industry newsletter

The latest AI trends, brought to you by experts

Get curated insights on the most important—and intriguing—AI news. Subscribe to our weekly Think newsletter. See the IBM Privacy Statement.
We use your email to validate you are who you say you are, to create your IBMid, and to contact you for account related matters.
Business email

Your subscription will be delivered in English. You will find an unsubscribe link in every newsletter. You can manage your subscriptions or unsubscribe here. Refer to our IBM Privacy Statement for more information.

Your subscription will be delivered in English. You will find an unsubscribe link in every newsletter. You can manage your subscriptions or unsubscribe here. Refer to our IBM Privacy Statement for more information.

Types of agentic memory

Researchers categorize agentic memory in much the same way that psychologists categorize human memory. The influential Cognitive Architectures for Language Agents (CoALA) paper[1] from a team at

Princeton University describes different types of memory as:

Short-term memory

Short-term memory (STM) enables an AI agent to remember recent inputs for immediate decision-making. This type of memory is useful in conversational AI, where maintaining context across multiple exchanges is required.

For example, a chatbot that remembers previous messages within a session can provide coherent responses instead of treating each user input in isolation, improving user experience. For example, OpenAI's ChatGPT retains chat history within a single session, helping to ensure smoother and more context-aware conversations.

STM is typically implemented using a rolling buffer or a context window, which holds a limited amount of recent data before being overwritten. While this approach improves continuity in short interactions, it does not retain information beyond the session, making it unsuitable for long-term personalization or learning.

Long-term memory

Long-term memory (LTM) allows AI agents to store and recall information across different sessions, making them more personalized and intelligent over time.

Unlike short-term memory, LTM is designed for permanent storage, often implemented using databases, knowledge graphs or vector embeddings. This type of memory is crucial for AI applications that require historical knowledge, such as personalized assistants and recommendation systems.

For example, an AI-powered customer support agent can remember previous interactions with a user and tailor responses accordingly, improving the overall customer experience.

One of the most effective techniques for implementing LTM is retrieval augmented generation (RAG), where the agent fetches relevant information from a stored knowledge base to enhance its responses.

Episodic memory

Episodic memory allows AI agents to recall specific past experiences, similar to how humans remember individual events. This type of memory is useful for case-based reasoning, where an AI learns from past events to make better decisions in the future.

Episodic memory is often implemented by logging key events, actions and their outcomes in a structured format that the agent can access when making decisions.

For example, an AI-powered financial advisor might remember a user's past investment choices and use that history to provide better recommendations. This memory type is also essential in robotics and autonomous systems, where an agent must recall past actions to navigate efficiently.

Semantic memory

Semantic memory is responsible for storing structured factual knowledge that an AI agent can retrieve and use for reasoning. Unlike episodic memory, which deals with specific events, semantic memory contains generalized information such as facts, definitions and rules.

AI agents typically implement semantic memory using knowledge bases, symbolic AI or vector embeddings, allowing them to process and retrieve relevant information efficiently. This type of memory is used in real-world applications that require domain expertise, such as legal AI assistants, medical diagnostic tools and enterprise knowledge management systems.

For example, an AI legal assistant can use its knowledge base to retrieve case precedents and provide accurate legal advice.

Procedural memory

Procedural memory in AI agents refers to the ability to store and recall skills, rules and learned behaviors that enable an agent to perform tasks automatically without explicit reasoning each time.

It is inspired by human procedural memory, which allows people to perform actions such as riding a bike or typing without consciously thinking about each step. In AI, procedural memory helps agents improve efficiency by automating complex sequences of actions based on prior experiences.

AI agents learn sequences of actions through training, often using reinforcement learning to optimize performance over time. By storing task-related procedures, AI agents can reduce computation time and respond faster to specific tasks without reprocessing data from scratch.