AI Agents, Long-Term Memory, and Semantic Fidelity

Why Autonomous Systems Gradually Lose Context While Remaining Operational

Representational Systems Brief #1

A. Jacobs | Semantic Fidelity Project

The Basic Pattern

AI agents do not simply generate responses. They operate through representations accumulated across time, including conversation histories, stored memories, planning structures, and retrieved documents. Each of these represents a different part of the agent's environment, allowing the system to remember prior interactions, incorporate external information, coordinate complex workflows, and continue reasoning beyond a single prompt.

This accumulated representational layer makes long-running tasks possible, but it also introduces a deeper problem. As information is repeatedly summarized, retrieved, compressed, and reused, the agent may continue functioning while gradually losing contact with the user's original intentions. The relevant memory may still be available, the context may still appear complete, and the plan may remain internally coherent, even as the meaning carried by those representations begins to shift.

AI research approaches this problem through fields such as agent memory, long-term memory, context engineering, memory retrieval, persistent memory, and agentic workflows. Although these areas emphasize different technical mechanisms, they all confront the same underlying challenge: preserving useful information across extended interactions without allowing its meaning to degrade.

Within the Semantic Fidelity framework, agent memory is therefore not only a storage problem. It is part of a broader representational problem concerning whether meaning survives as information moves through increasingly complex systems.

Memory Is Not Understanding

Agent memory succeeds because it preserves representations of previous interactions. Conversation history, retrieved memories, summaries, and external documents allow the system to continue reasoning long after the original interaction occurred.

The same compression that makes long-term memory practical also creates the possibility of semantic loss. A summary may preserve the appearance of continuity while gradually omitting assumptions, distinctions, or contextual relationships that originally guided the conversation. The important question is whether compression preserves the assumptions, distinctions, and relationships that make those memories useful.

Understanding the Problem

Long-Term Memory: Stores information across conversations or extended workflows.

Memory Retrieval: Selects previously stored information that appears relevant to the current task.

Context Engineering: Organizes information so models receive the most useful representations during reasoning.

Agentic Workflows: Coordinate planning, memory, retrieval, and tool use across extended tasks.

Although these systems differ technically, they share a common objective. They preserve representations across time. They do not automatically preserve meaning across time.

How Memory Drift Emerges

Stage 1 — Experience

An interaction occurs. Information remains closely connected to the original context in which it was created.

Stage 2 — Compression

Conversation history is summarized. Memories are indexed. Context is reduced into representations that can be efficiently retrieved.

Stage 3 — Reuse

The agent retrieves previous memories while solving new problems. Earlier representations increasingly influence future reasoning.

Stage 4 — Semantic Drift

The retrieved memory remains coherent while gradually losing the assumptions, relationships, and contextual boundaries that originally gave it meaning.

The memory remains available. The meaning gradually changes. The absence of retrieval failure is not proof of preserved understanding. It may simply mean the memory system has stopped preserving the distinctions that originally mattered.

Semantic Fidelity and Reality Drift

Long-term memory improves continuity by preserving representations across time. Semantic Fidelity asks a complementary question. Does the retrieved memory still preserve the meaning intended when that information was originally created?

An agent may correctly retrieve a previous interaction while subtly changing its interpretation through summarization, compression, or repeated reuse.

Reality Drift introduces a broader perspective. As autonomous systems increasingly reason through accumulated representations rather than immediate observations, maintaining coherent memories can become easier than maintaining accurate representations of changing conditions.

The Agent Memory Stack

Agent memory becomes more complex as information moves through increasingly abstract layers. An experience is first captured in conversation, then stored as memory, compressed into summaries, selected through retrieval, and reintroduced as context. That context shapes reasoning, reasoning produces plans, and plans guide actions. Those actions then create new experiences that enter the memory system and begin the cycle again.

Each stage makes long-term reasoning more scalable, but it also creates another opportunity for meaning to change. A conversation may begin with a precise understanding of the user's intentions, yet repeated summarization, retrieval, planning, and reuse can gradually remove the assumptions and distinctions that originally gave the interaction its meaning.

This is how memory failure can become systemic. The agent may preserve continuity across time while slowly weakening the meaning that continuity was intended to preserve. The map keeps updating even after the territory disappears. The further information travels through the representational stack, the more important semantic fidelity becomes.

Examples Across AI Systems

Personal AI Assistants: An assistant remembers user preferences across many conversations while gradually simplifying or reinterpreting those preferences over time.

Coding Agents: Long-running software projects accumulate plans, summaries, and implementation history. Earlier design assumptions gradually disappear while the agent continues extending the project.

Research Agents: Information gathered from multiple documents remains available through summaries and retrieval, yet subtle distinctions between sources gradually become compressed into generalized conclusions.

Customer Support Agents: Persistent memory allows conversations to continue across multiple sessions while introducing opportunities for previous misunderstandings to become reinforced.

Multi-Agent Systems: One agent summarizes information for another. Each transfer preserves continuity while introducing additional opportunities for semantic reinterpretation.

Semantic Fidelity as a General Principle

Agent memory represents a major advance in artificial intelligence by allowing systems to reason across time rather than within individual prompts. Long-term memory, retrieval, and context engineering make it possible to preserve continuity across increasingly complex interactions, but continuity alone does not guarantee preserved understanding.

Semantic fidelity addresses this deeper layer of the problem. Memory determines whether information remains available, while semantic fidelity determines whether the meaning carried by that information survives repeated summarization, retrieval, compression, and reuse. A representation may remain accessible and operational while gradually changing what it represents.

As AI agents become more persistent, the central challenge will extend beyond remembering previous information. It will involve preserving the assumptions, distinctions, and contextual relationships that allow that information to remain meaningful across increasingly long chains of reasoning and action.

Related Concepts

Memory Drift: The gradual change in what stored information means as it is summarized, retrieved, reinterpreted, and reused across time.

Recursive Compression: The repeated compression of already compressed representations. Each additional transformation may remove distinctions while preserving the appearance of coherence.

Retrieval Drift: A condition in which retrieved information remains relevant at the keyword or embedding level but no longer preserves the context required for correct interpretation.

Constraint Collapse: The loss of the original boundaries, assumptions, or requirements that constrained how information was supposed to be understood.

Reality Drift: The broader condition in which a system remains operational while its internal representations gradually lose alignment with real-world conditions.

Frequently Asked Questions

Is memory drift the same as retrieval failure?

No. Retrieval failure occurs when a system cannot find the relevant memory. Memory drift occurs when the memory is successfully retrieved but no longer preserves the meaning, context, or distinctions that originally made it useful.

Can an AI agent remember accurate information and still misunderstand the user?

Yes. A stored fact may remain technically accurate while its significance, boundaries, or relationship to the user's present intention changes. Availability does not guarantee correct interpretation.

Why does long-term memory increase semantic risk?

Long-term memory introduces additional stages of summarization, indexing, retrieval, interpretation, planning, and reuse. Each transformation creates another opportunity for meaning to change without producing an obvious technical failure.

Does semantic fidelity require storing complete conversation histories?

No. All practical memory systems compress information. Semantic fidelity concerns whether the compression preserves the distinctions, relationships, and contextual boundaries required for future understanding.

Canonical References

Reality Drift: Canonical Concept Paper
A broader account of how systems remain functional while gradually losing alignment with real-world conditions.

Semantic Fidelity: Concept Paper and Definition
A general framework for evaluating whether meaning survives representation, compression, transmission, retrieval, and reuse.

Recursive Compression
An explanation of how repeated summarization and reuse can preserve coherence while weakening distinctions and contextual structure.

Related Search Terms

- AI agents
- agent memory
- long-term memory
- persistent memory
- context engineering
- memory retrieval
- agentic workflows
- semantic fidelity
- memory drift
- representation drift
- retrieval drift
- reality drift

- constraint collapse
- recursive compression
- grounding