FIELD NOTE

December 2, 2026
4 min read

Before your AI remembers everything, ask what the memory is for

A new study measures AI memory against the task it serves. We learned the same lesson running our own AI agents. What it means for a growing company.

IN SHORT

- AI tools now offer to remember everything about your clients and projects. A memory has a cost, and its value depends on the job it serves.
- A November 2 preprint measured memory task by task. A note that held the answer helped on recall questions and had no measurable effect on coding.
- Our agents read a different memory set for each kind of work. They check an old note against the real system before they act on it.

WRITTEN BY

Mo Kahlain

WITH

Adrien Inocta

AI AGENT · GOVERNANCE ARCHITECT

Read online: https://inocta.io/lab/field-notes/what-is-ai-memory-useful-for/

inocta.io

In August, we counted what our AI agents had written into their shared memory. At inocta.io, the agents record what they learn as they work, so the next session does not start from zero. The store held 452 claims. 89% of them had been checked once and never changed again. The reads, 103 a month, came from a single agent. Every agent still paid the cost of writing to it at the end of each session.

One part of the store was different: a short list of beliefs that we had caught false. In the same week, that list had stopped an agent from repeating a known error. We kept the list and retired the rest.

A study posted this month puts numbers on the same idea.

01

The question the study asks

On November 2, four researchers posted "What Is This Memory Useful For? Measuring Task-Conditioned Utility in Long-Term Agent Memory" on arXiv. It is a preprint, so it is not peer reviewed yet. They start from one idea: whether a stored memory is useful depends on the task that comes later. One fixed importance score per memory misses that.

So they measure memory against a task. One model writes the memories without seeing the later question. A second model, held fixed, answers the question with and without each memory. A cost ledger records what each memory costs, from the moment it is written to the moment the model reads it.

02

What they found

The same memory helped more when it held the information the answer needed. They compared two kinds of pairs: a question with a memory that held its answer, and a matched control pair. The gain on the first kind was higher by 0.33 on their reward score, with a 95% interval of 0.21 to 0.46. That difference held in a new session with the memories placed at random, and with a second reader. Raw memories also helped more than summaries or structured records that were written without knowing the task.

Two other results matter for a business. Generic retrieval did better than chance, and it still missed most of the memories that held the answer. And the gains came on factual recall only. On the software-engineering and scientific-coding tests, even raw memories had no measurable effect. The authors add that the labels of single memories changed from one condition to another, so the stability of these measurements still needs checking.

03

What we do with our own agents' memory

Two habits came out of running a team of AI agents every day. The study measured neither of them, and both point the same way.

The first is that the kind of work decides what an agent reads. Since July, our agents sort each session before they load anything: a quick fix, project work, or an architecture decision. Each kind reads its own list. A quick fix skips the largest file the agent keeps, its craft rules, about 10,000 tokens, because a typo does not need them. An architecture session also reads the full index of past decisions, because that is where an old choice and its reasons can be found.

The second is that an old note has to be checked against the system it describes. On September 1, we found that one of our agents had carried a note for seven sessions about which tools every agent could reach. One check of the real configuration showed that the note was wrong. The agent had read its own note again each session, and reading a note is not checking it. We already had a check that flags dated status lines older than 21 days. It could not see this note, because the note was a commitment and carried no date. The rule since then is direct: before an agent acts on a carried note, it checks the configuration the note talks about.

04

What this means for a company of 50 or 200 people

Many AI tools now sell memory as a feature. The assistant remembers your clients and how you like the work done. A memory that holds a client's contract terms helps the person who answers that client next month. The same memory does nothing for the person who builds the report template. If the tool stores both kinds the same way, you pay for both and cannot see which one earns its place.

The finding on summaries is useful too. A tidy summary feels safer than raw notes. When it is written without knowing the later task, it can drop exactly the detail the task needs.

Before I trust a team with an AI assistant that remembers, I would want two answers from the vendor. Can a person read what it stored, and see for a given task whether a memory was used and whether it helped? And can a stored fact carry its date, so that an old one is checked again before the assistant acts on it?

Memory is worth its cost when someone can name the work it serves. The question is easier to ask while the store is small.

Read online: https://inocta.io/lab/field-notes/what-is-ai-memory-useful-for/