Why AI forgets you between conversations, and what to do about it

How the Models Work

2 September 2026

Malte Clausen
malteclausen.com

Why AI forgets you between conversations, and what to do about it

AI gives you a conversation partner with all the world's knowledge and no memory of you.

Think of a colleague you talk to every week. Every conversation gives you three things at once:

1. He knows things about the world.
2. He knows things about you.
3. Each conversation builds on the ones before.

You get all three automatically. That is just how talking to a person works.

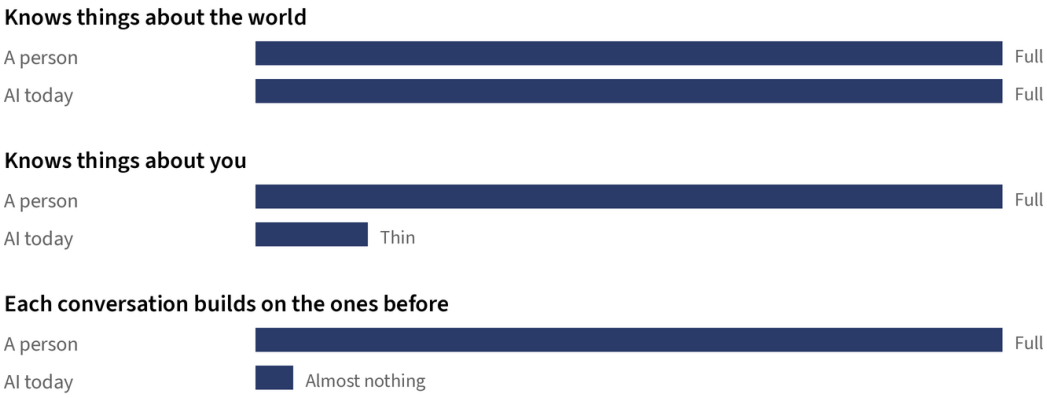
AI is the first conversation partner that splits them apart. The first thing, knowledge of the world, you get in abundance: the model has read more than any human ever will. The second thing, memory of you, is thin. AI tools like ChatGPT and Gemini keep short notes about you: a few lines of facts and preferences. Those notes are what "remembering" means. The third thing barely exists. Each new conversation starts blank, whatever happened yesterday.

This is why talking to AI feels the way it does. You expected a brilliant colleague. You got a brilliant specialist you have never met before, every single time.

The picture that fits is the doctor and the patient file. A specialist knows medicine completely. Years of training, thousands of cases, all of it present in every consultation. But he does not know you. Everything about your situation comes from the file on the desk: the record of every earlier visit. If your last visit is missing from the file, that visit never happened, as far as he is concerned. AI works the same way. The model learned its general knowledge during training, the way the doctor learned medicine in school. The file is whatever you place in front of it today. And the memory features these tools now advertise are notes in the file: two lines written after each visit, useful and thin.

Why not simply hand the model the full transcript of everything you have ever discussed? Because the model re-reads its entire file for every single answer, and the more it reads, the less attention each fact gets. I wrote about that effect in the context window piece. Full recall does not scale. The notes exist because the file must stay short.

What a human partner gives you, and what AI gives you



Illustrative comparison, not a measurement.

Three habits follow:

1. End every working session by asking the AI to write the summary. One line does it: "summarize the key decisions and status from this conversation."
2. Keep the summaries in one document that you own. That document is your file.
3. Start a new conversation by handing over the file. Paste the document in, or point the AI to it, before you ask for anything.

If that pattern sounds familiar, it is the same pattern my RAG piece describes, done by hand: retrieve what matters, place it in front of the model, then ask.

People are not wrong to expect all three things. The AI companies are racing to build the third one, so that one day every conversation will build on the ones before. Until then, the quiet lesson: the relationship lives in the file, so keep the file.

Where this comes from

Two research papers sit behind this piece.

Researchers at Stanford measured what happens when a model must find one fact buried deep in a long input. The model often misses it. Their paper is called Lost in the Middle, and it is the research behind the line "the more it reads, the less attention each fact gets."

Researchers at Berkeley built a system that gives a model the same routine this piece recommends: short notes always in front of it, and a larger file it opens when needed. Their paper is called MemGPT. Keeping one document and handing it over is the same idea, done by hand.