

AI forgets. Why do I have to keep repeating myself?

A new chat is not always a blank slate. Knowing the difference is what stops the repeating.

thirdbrain.

Start →

1 A new chat starts with a new transcript

The messages in one chat are not automatically copied into another. That is useful when the jobs are unrelated. Memory changes the picture:

Claude, ChatGPT and Gemini can carry selected information from previous conversations when memory is enabled. A project can also provide its own instructions and files to every chat inside it.

2 What the model can see

The model answers from the context available for that request. This can include your instructions, relevant memories, project files, connected sources, the messages in this chat and the message you just sent.

If important information is missing, it falls back to general knowledge from training or web search when that feature is available. That may not match your situation.

3 Every message builds on the last

Inside one chat, each message becomes part of the conversation the model uses for the next response.

A chat used for several unrelated jobs can become noisy. Old assumptions and examples stay in the transcript and can pull a later answer in the wrong direction.

One job, one chat.

4 The context window is the working space

The context window is the amount of material a model can work with at once. The limit depends on the tool, model, plan and feature you are using.

A token is a small piece of text that the model counts as it reads and writes.

According to Anthropic, current paid Claude plans offer context windows from 200,000 to 1 million tokens, depending on the model. That is roughly 500 to 2,500 pages of plain text, but images, files, tool output and the model's response also use the available space.

A large limit is not a reason to fill one chat with everything. Relevant context beats maximum context.

5 What happens when the context fills up

When the context window fills, the tool usually compacts older material, warns you about a length limit, or asks you to start a new chat.

Compaction is useful, but a summary cannot preserve every detail. If accuracy starts slipping, open a fresh chat and bring across only the context the next step needs.

6 How to stop repeating yourself

Stop repeating yourself by putting stable facts in instructions, reusable files in a project, and selected details in memory.

INSTRUCTIONS

Write stable facts and standing preferences once.

Use Australian English and give me the recommendation first.

PROJECTS

Keep the files, instructions and conversations for one body of work together.

A hiring project with the role brief, interview notes and decision criteria.

MEMORY

Let the tool retain selected information across conversations, then review and correct it.

My role, recurring preferences or a long-running goal.

Memory availability and defaults vary by product, plan and organisation policy. Microsoft Copilot calls its version of a project a notebook. On Copilot, whether the tool can reach your actual email and files depends on the licence your organisation bought, so check before you assume it can see your work.

7 Try this in two minutes

Open the memory settings, read what the tool may use about you, then correct or remove anything that is wrong.

01 Open the memory settings.

Claude: Settings > Memory. ChatGPT: Settings > Personalization > Memory. Gemini: Settings & help > Personal Intelligence > Memory.

02 Read what the tool may use about you.

03 Correct anything that is wrong or out of date.

04 Remove anything you do not want carried into future chats.

Are any of the memories wrong?

A FIRST CONVERSATION

Want help applying this to your work?

Bring one real task. We will work through it together and sharpen your setup as we go. No deck, no pitch.

thirdbrain.com.au

hello@thirdbrain.com.au

thirdbrain.