

EVERYONE · 5 MINUTE READ

How AI assistants work, and where they fail

AI assistants predict likely text from patterns, so they write fluently but can be confidently wrong; use them for drafts and check the facts.

What it is

An AI assistant, such as Claude, ChatGPT, Microsoft Copilot or Gemini, is built on a large language model. A language model is a computer program trained on a huge amount of text. From that text it learned patterns: which words, ideas and styles tend to follow each other. When you type a question, it writes an answer one small piece at a time, each time choosing a likely next piece.

Think of a very well-read colleague who answers instantly from memory and rarely says "I'm not sure". Usually the answer is useful; sometimes it is just as confident and wrong. The assistant is not looking things up in a database of facts unless it has been given a tool to do so, such as web search or access to your documents.

So it is strong with language: drafting, rewording, summarising, sorting and explaining. It is weaker where exactness matters: precise figures, dates, quotes, sums, and anything about your organisation it has never seen.

Why it matters

- It can produce **hallucinations**: answers that look correct but are invented, such as a policy that does not exist or a reference to a report nobody wrote.
- It has a **knowledge cut-off**: a date after which it learned nothing. Without web search it will not know recent news, prices or rule changes.
- A new chat may or may not see earlier chats: some tools have memory features (sometimes on by default) or projects that carry details over. Check your settings, and do not assume it remembers or forgets.
- It does not know your customers, your team's habits or what "urgent" means in your department, unless you tell it.

How to do it

1. Use the assistant for language tasks first: first drafts, summaries, rewording, lists of options.
2. Give it the background it cannot know, such as who the reader is and what the document is for.
3. Ask for a draft, not a final answer, and expect to edit it.
4. Check every fact, number, name and date against a source you trust before you use it.
5. Use a calculator or spreadsheet for sums, or ask the assistant to use its calculation tool if it has one, then check the result.
6. Turn on web search, or paste in the document, when the answer depends on recent or specific information.
7. Start a fresh chat when you change topic, so old details do not mix in.

Example

Before: "What is our refund policy for business customers?"

The assistant has never seen your policy, so it may write one that sounds official and is completely made up.

After: paste in the approved policy text (if your organisation allows it), then ask:

Below is our refund policy. Summarise it in five bullet points for a new customer service colleague. Use only what is in the text. If something is not covered, write "not covered".

[paste approved policy text]

Now the assistant is working from your material, not guessing. A person still reads the summary against the policy before sharing it.

Common mistakes

- **Treating a fluent answer as a true answer.** Fix: confidence in the writing says nothing about accuracy, so check facts.
- **Assuming it remembers last week's chat.** Fix: paste the key background again, or use a project or saved instructions if your tool has them.
- **Asking about recent events without web search.** Fix: switch on search or give it the source, and look at the links it cites.
- **Trusting its arithmetic.** Fix: check totals in a spreadsheet.
- **Pasting confidential data into an unapproved tool.** Fix: use only tools your organisation has approved, and start with made-up examples.

Try it this week

1. Pick a topic you know very well from your own job.
2. Ask the assistant three questions about it: one general, one with a specific number, one about something recent.
3. Mark each answer: correct, partly correct, or wrong.
4. For any wrong answer, ask "How sure are you, and what is your source?" and see how it responds.
5. Write one sentence for yourself: "I will trust it for ___ and always check ___."

Going further

- Memory features and saved instructions change what the assistant "knows" about you; review what is stored from time to time.
- Asking the same question twice can give different answers, because the model picks from likely options. Differences are a useful warning sign.
- When an assistant is connected to documents or tools, it can quote from them, but it can still misread them. Check the quote against the original.

Key words

- **Language model:** a program trained on large amounts of text to predict likely next words.
- **Hallucination:** an answer that sounds right but is invented or wrong.
- **Knowledge cut-off:** the date after which the model learned nothing new.
- **Context:** the information the assistant has in front of it in this chat.
- **Memory feature:** a setting that lets an assistant keep some details between chats.