What is a context window in a large language model?

1 November 2026 · 4 min read · blogs.iictp.org/what-is-an-llm-context-window

A context window is the maximum amount of text, measured in tokens, that a language model can consider in one request. It covers everything: your instructions, the conversation so far, any documents you supply and the model's own reply.

TL;DR The short version

- A context window is the most text, in tokens, a model can consider in one request, including instructions, conversation, documents and its reply.
- A bigger window does not automatically mean better answers, and long inputs cost more and take longer.
- Work within it by retrieving only relevant passages, summarising earlier turns and splitting large documents.

Read online, listen to it, or change the text size: https://blogs.iictp.org/what-is-an-llm-context-window

What is a token?

A token is a piece of text the model works with, often a word fragment. Common English words are often one token, while rarer words, code and many non-English languages are split into more tokens. Limits and prices are usually counted in tokens, not words.

What happens when the window fills up?

Either the request is rejected, or older content is dropped or summarised by the application. In a long chat, early instructions can be pushed out, which is why a model may seem to forget rules you gave at the start.

Does a bigger window mean better answers?

Not automatically. Models can use information in the middle of a very long input less reliably than at the start or end, and more tokens mean higher cost and slower responses. Sending only the relevant material usually works better than sending everything.

How do you work within the limit?

Retrieve only the passages that matter, summarise earlier turns, split large documents into parts and combine the results, and keep standing instructions short and near the start. Restate key rules when a conversation runs long.

How does it affect cost and speed?

Providers generally charge per token for both input and output, and long inputs take longer to process. Reusing a stable prefix across requests, where the provider supports caching, can reduce both.

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Liu et al. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172.
2. Vaswani et al. Attention Is All You Need. arXiv:1706.03762.

Published by IICTP, an independent technical credential body, on 1 November 2026. The current version of this post is at https://blogs.iictp.org/what-is-an-llm-context-window. More guides: blogs.iictp.org.