Tokenization & Context Windows

> **Plain English:** AI doesn't read words — it reads "tokens" (word chunks). Every AI tool has a limit on how many tokens it can handle at once. This is the context window, and it's the single biggest constraint on what AI can do for you.

What tokens actually are

When you type a message to AI, it doesn't see words the way you do. It breaks your text into smaller pieces called **tokens**. Sometimes a token is a whole word. Sometimes it's a piece of a word. Sometimes it's just punctuation.

For example:

- "Hello" → 1 token
- "Tokenization" → might become "Token" + "ization" → 2 tokens
- "I'm building a workflow" → roughly 5 tokens

A rough rule of thumb: **1 token ≈ ¾ of a word** in English. So 1,000 words is roughly 1,300 tokens. A 10-page document is roughly 4,000–5,000 tokens.

Why does this matter to you? Because every AI tool charges by tokens and limits by tokens. When you hit a message length limit, get a "conversation too long" error, or notice AI "forgetting" things you said earlier — that's all about tokens.

The context window: AI's working memory

The **context window** is the total number of tokens the AI can process at once. Think of it as a whiteboard: everything in your conversation — your messages, the AI's responses, any documents you've uploaded, the system prompt — all has to fit on this whiteboard. When it's full, things start falling off the other end.

Current context window sizes (as of early 2026):

Model	Context window	Roughly equivalent to
Claude (Anthropic)	200K tokens	~150,000 words — a full novel
GPT-4o (OpenAI)	128K tokens	~96,000 words
Gemini 1.5 Pro (Google)	1M+ tokens	~750,000 words — multiple books

These numbers sound enormous, but they fill up faster than you'd think. A long conversation with back-and-forth responses, a few uploaded documents, and a detailed system prompt can eat through 200K tokens in a working session.

Why this matters for your work

Long conversations degrade. If you've noticed AI giving worse answers later in a conversation than at the beginning, this is why. As the context window fills up, the model has more text to process and older information gets less "attention." Starting a fresh conversation for a new topic isn't a sign of failure — it's good practice.

Uploaded documents have limits. When you upload a PDF or paste a long document, it consumes context window space. A 50-page report might use 20,000+ tokens, leaving less room for your actual questions and the AI's responses. If you're working with long documents, consider summarizing or extracting the relevant sections first.

"AI forgot what I said" is usually a context issue. AI doesn't have memory between conversations (unless you're using features like Claude's Projects or custom GPTs that provide persistent context). Even within a conversation, if you're 30 messages in, the AI may lose track of something you said at the beginning because it's being pushed out of the active window.

This is why the Handoff Protocol exercise matters. When you learn to structure handoffs between AI sessions — summarizing context, carrying forward the essential information — you're working around context window limits intelligently.

Practical tips

Start fresh for new topics. Don't keep one mega-conversation running for everything. A new topic deserves a new conversation with focused context.

Front-load the important stuff. Put your most critical instructions, context, and constraints at the beginning of your prompt. Information at the start and end of the context window gets more "attention" than information buried in the middle.

Summarize before you continue. If a conversation is getting long and you want to keep going, ask the AI to summarize the key decisions and context so far, then start a new conversation with

that summary.

Use persistent context features. Claude Projects, custom GPTs, and system prompts let you set context that persists across messages without eating into your per-message token budget. The Handoff Protocol exercise teaches you to design these.

Be selective with document uploads. Instead of uploading a 100-page document and asking a question, extract the 5 relevant pages. You'll get better answers and use less of your context budget.

Where to go next

- Prompt Engineering Basics — how to make the most of limited context
- The Handoff Protocol — practice managing context across AI sessions

---

Revision #1
Created 2026-03-16 15:47:57 UTC by Admin
Updated 2026-03-16 15:47:57 UTC by Admin