Claude Context Window: 1M Tokens for Opus 5.5, Sonnet 5.5 & Fable 5.1 (2026)

Edition of November 13, 2026 — facts verified November 13, 2026. Current edition always at https://tygartmedia.com/claude-context-window/.

Direct Answer (November 13, 2026): A **1M-token context window** (~750,000 words / ~3,000 pages) is now standard across **Fable 5.2, Opus 5.5, Sonnet 5.5, and Haiku 5.5**. Legacy Haiku 4.5 uses a **200K-token window** (~150,000 words / ~500 pages), which remains a useful mental benchmark for lighter prompts.

Last refreshed: November 13, 2026 (Pacific)

Model Accuracy Note — Updated November 13, 2026

Current lineup (per Anthropic models overview & API pricing): **Claude Fable 5.2** (`claude-fable-5-2`, $8/$40), **Claude Opus 5.5** (`claude-opus-5-5`, $4/$20), **Sonnet 5.5** ($2/$10), and **Haiku 5.5** ($0.10/$0.50 ≤100K) — all with a **1M context window** and 128K max output. Legacy Haiku 4.5 ($1/$5) remains 200K/64K. See current model tracker →

Claude AI · Fitted Claude

Claude's context window determines how much information it can hold and process in a single conversation. All four current models — **Claude Fable 5.2, Opus 5.5, Sonnet 5.5, and Haiku 5.5** — support a **1 million token** context window (128K max output) at standard per-token pricing, while legacy Haiku 4.5 remains at 200,000 tokens.

Claude Context Window by Model (November 2026)

Claude context window by model.

Model	Context Window	~Words	~Pages
Claude Opus 5.5	1,000,000 tokens	~750,000	~3,000
Claude Sonnet 5.5	1,000,000 tokens	~750,000	~3,000
Claude Fable 5.2	1,000,000 tokens	~750,000	~3,000
Claude Haiku 5.5	1,000,000 tokens	~750,000	~3,000
Claude Haiku 4.5 (legacy)	200,000 tokens	~150,000	~500

What Fits in 200K Tokens

What fits in 200K tokens.

Content type	Approximate fit
News articles	~200+ articles
Research papers	~30–50 papers depending on length
A full novel	Yes — most novels fit with room to spare
Python codebase	Medium-sized codebases (10k–50k lines)
Legal contracts	Hundreds of pages of contracts
Conversation history	Very long sessions before truncation

Context Window vs. Output Length

The context window covers everything Claude processes — both input and output combined. If your prompt is 50,000 tokens (a long document), Claude has 150,000 tokens remaining for its response and any further back-and-forth. The window is shared between what you send and what Claude generates.

Maximum output length is a separate constraint — Claude won't generate an infinitely long response even within a large context window. For very long outputs (full books, extensive reports), you typically work in sections rather than expecting Claude to produce everything in one pass.

Why Context Window Size Matters

Why context window size matters.

Context window size is the practical limit on how much work you can give Claude at once without losing information. Before large context windows, working with long documents required chunking — splitting the document into pieces, analyzing each separately, and manually synthesizing the results. With 200K tokens, Claude can hold the entire document and answer questions about any part of it with full awareness of everything else.

This matters most for: document analysis and legal review, code understanding across large files, research synthesis across many sources, and long multi-step conversations where earlier context affects later decisions.

How Claude Performs at the Edges of Its Context Window

Research on large language models has found that performance can degrade somewhat for information buried in the middle of a very long context — sometimes called the "lost in the middle" problem. Claude performs well across its context window, but for maximum reliability on information from a very long document, referencing specific sections explicitly ("in the section about pricing on page 12…") helps ensure Claude focuses on the right part.

For the full model spec breakdown, see Claude API Model Strings and Specs and Claude Models Explained: Haiku vs Sonnet vs Opus.

Frequently Asked Questions

What is Claude's context window size?

Claude Fable 5.2, Opus 5.5, Sonnet 5.5, and Haiku 5.5 all support a **1 million token** context window (~750,000 words / ~3,000 pages) and 128K output tokens. Legacy Claude Haiku 4.5 supports 200,000 tokens (~150,000 words / ~500 pages).

How many tokens is 200K context?

200,000 tokens is approximately 150,000 words of English text. One token is roughly four characters or three-quarters of a word. A typical 800-word article is about 1,000 tokens; a full novel is typically 80,000–120,000 tokens.

Can I upload a full PDF to Claude?

Yes — current models provide a 1M token context window (200K on legacy Haiku 4.5), and individual PDF uploads also respect product upload file-size and page-count limits. Multi-volume sets above the context or upload ceiling should still be split.

What is Claude's context window in 2026?

Fable 5.2, Opus 5.5, Sonnet 5.5, and Haiku 5.5 all feature a **1 million token** context window (~750,000 words) and 128K max output tokens. Legacy Haiku 4.5 retains a 200,000-token context window. The 1M context window is generally available across current models.

How many pages can Claude read in one context window?

At 1 million tokens, Claude Opus 5.5 can process approximately 750,000 words — roughly 3,000 pages of standard text, 10 full novels, or a large software codebase. This makes it practical for full repository analysis, entire book summarization, and extended agent sessions without needing to chunk input.

Does Claude charge extra for large context windows?

No. Anthropic eliminated long-context surcharges in March 2026. A 900,000-token API call is billed at the same per-token rate as a 9,000-token call. All tokens in the context window — input and output — are billed at the standard rates for

whichever model you are using.

Related:

Can Claude Read PDFs? Yes — Here’s How It Works

Related:

Claude Token Limit: Context Windows and Output Limits Explained

Need this set up for your team?

Talk to Will →

Check current windows any time from the Anthropic Console — API keys, billing, and a token-to-dollar cost estimator.

Primary & Reference Links Cited on This Page

- Anthropic models overview — `https://platform.claude.com/docs/en/about-claude/models/overview`
- API pricing — `https://platform.claude.com/docs/en/about-claude/pricing`
- Tygart Media: current model tracker →
- Tygart Media: Claude API Model Strings and Specs
- Tygart Media: Claude Models Explained: Haiku vs Sonnet vs Opus
- Tygart Media: Can Claude Read PDFs? Yes — Here’s How It Works
- Tygart Media: Claude Token Limit: Context Windows and Output Limits Explained
- Tygart Media: Talk to Will →
- Tygart Media: Anthropic Console