TYGART MEDIA · CLAUDE & ANTHROPIC REFERENCE

Claude Token Limit: Context Windows, Output Limits, and What They Mean in Practice

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

Direct Answer (November 13, 2026): All four current Claude models — **Fable 5.2, Opus 5.5, Sonnet 5.5, and Haiku 5.5** — share a **1,000,000-token context window (~750,000 words or ~1,500 pages)** with up to **128,000 tokens of max output** per response. Legacy **Haiku 4.5** (200,000-token context, 64,000-token max output) and legacy **Fable 5.1** remain on the overview.

Updated October 28, 2026: Made infographic diagrams clickable to open full-size images.

Last verified: November 13, 2026 (Pacific Time) · Anthropic models overview · API pricing

Model Accuracy Note — Updated November 13, 2026

Current models (November 13, 2026): **Fable 5.2 · Opus 5.5 · Sonnet 5.5 · Haiku 5.5** — all 1M context, 128K max output. Current model tracker →

Claude AI · Fitted Claude

Claude's token limits depend on which model you're using and whether you're on the web interface or the API. Here are the exact numbers — context window, output limits, and what they mean in practice.

Key distinction:

The context window is the total tokens Claude can process in one conversation (input + output combined). The output limit is the maximum tokens in a single response. These are different limits and both matter depending on your use case.

Claude Token Limits by Model (November 13, 2026)

Model	Context Window	Max Output
Claude Fable 5.2	1,000,000 tokens	128,000 tokens
Claude Opus 5.5	1,000,000 tokens	128,000 tokens
Claude Sonnet 5.5	1,000,000 tokens	128,000 tokens
Claude Haiku 5.5	1,000,000 tokens	128,000 tokens

Context windows are at standard API pricing (no long-context premium) on most models; Haiku 5.5 is the exception, stepping from $0.10/$0.50 to $0.50/$2.50 per MTok on prompts above 100k tokens. Claude Opus 5 (released July 24, 2026) is superseded by Opus 5.5 (released September 22, 2026) at a lower price ($4/$20 vs $5/$25 per MTok). For new projects, use Opus 5.5 — the recommended default for most workloads. Outputs above ~16,000 tokens should be streamed to avoid request timeouts. The original Opus 4 and Sonnet 4 were retired on June 15, 2026 — see the deprecation guide. The asynchronous Message Batches API runs at 50% cost and supports larger per-request outputs for document-generation workloads.

What a Token Is

Count tokens when planning long docs — not characters alone.

A token is roughly 3–4 characters of English text — about 0.75 words. One page of text is approximately 500–700 tokens. A 200-page book is roughly 100,000–140,000 tokens.

Content	Approx. tokens
1 word	~1.3 tokens

Content	Approx. tokens
1 page of text (~500 words)	~650 tokens
Short novel (80,000 words)	~104,000 tokens
Full codebase (10,000 lines)	~100,000–200,000 tokens
1M token context (Sonnet/Opus)	~750,000 words / ~1,500 pages

Context Window vs. Output Limit

Two different ceilings. Hitting one does not mean you hit the other.

The context window is the total working memory for a session — everything Claude can "see" at once, including the system prompt, all previous messages in the conversation, uploaded files, and Claude's own prior responses. At 1M tokens, the legacy (still listed) models (Fable 5, Opus 4.8 *(legacy — still listed)*, Sonnet 4.6 *(legacy — still listed)*) can hold roughly 1,500 pages of text in context simultaneously.

The output limit is how long Claude's individual response can be, and it is model-dependent: up to 128,000 tokens on current Claude Fable 5.2, Opus 5.5, Sonnet 5.5, and Haiku 5.5 (64,000 tokens on legacy Haiku 4.5). Outputs above roughly 16,000 tokens should be streamed to avoid request timeouts.

Lineup currency (November 13, 2026): Current API list: Fable 5.2 $8/$40, Opus 5.5 $4/$20, Sonnet 5.5 $2/$10, Haiku 5.5 $0.10/$0.50 (≤100k tokens; $0.50/$2.50 >100k). Legacy (still listed): Fable 5.1 $10/$50, Haiku 4.5 $1/$5, Opus 4.8 $5/$25, Sonnet 4.6 $3/$15 (Anthropic API pricing).

Rate Limits: Separate From Token Limits

Token limits are per-conversation. Rate limits are per-time-period — how many tokens (and requests) you can send across multiple conversations in a given minute or day. Rate limits scale with your API usage tier. If you're hitting errors in production that look like limits, check whether you're hitting the context window, the output limit, or a rate limit — they produce different error codes. For the full rate limit breakdown, see Claude Rate Limits: What They Are and How to Work Around Them.

What Happens When You Hit the Context Limit

Soft edge: summarize, start a fresh thread, or drop older turns.

In claude.ai conversations, you'll see a warning when the conversation is approaching the context window. Claude may summarize earlier parts of the conversation to stay within limits. In the API, sending more tokens than the context window allows returns an error. For very long sessions, breaking work into multiple conversations or using prompt caching (which stores static context at a discount) are the standard approaches.

Frequently Asked Questions

What is Claude's token limit?

Claude Fable 5.2, Opus 5.5, Sonnet 5.5, and Haiku 5.5 all share a 1 million token context window. Legacy Claude Haiku 4.5 has a 200,000 token context window. Maximum output per response is 128,000 tokens on all four current models. These are different limits — context window is total working memory, output limit is maximum response length.

How long can Claude's responses be?

All four current models — Fable 5.2, Opus 5.5, Sonnet 5.5, and Haiku 5.5 — support up to **128,000 tokens of output** (~96,000 words). Older legacy models Sonnet 4.6 and Haiku 4.5 capped at 64,000 tokens. Outputs above ~16,000 tokens should be streamed. In practice, claude.ai conversations have shorter limits than the raw API.

How many tokens is a page of text?

Approximately 650 tokens per page (roughly 500 words). A 200-page document is around 130,000 tokens — well within the 1M-token context window shared by Fable 5.2, Opus 5.5, Sonnet 5.5, and Haiku 5.5 (and within legacy Haiku 4.5's 200K

window).

How many tokens can Claude Opus 4.8 handle?

Claude Opus 4.8 supports a 1 million token context window for input and up to 128,000 tokens for output. This means you can feed in very large documents, entire codebases, or long conversation histories in a single request.

What happens when you hit Claude's context window limit?

When your input exceeds Claude's context window, the API returns an error and the request fails. In claude.ai, the interface warns you before sending. The fix is to summarize earlier conversation turns, split the input into chunks, or use a model with a larger context window.

Is Claude Haiku's context window smaller than Opus?

Not on the current generation: **Claude Haiku 5.5** now has the full **1 million token context window** (and 128K output), matching Fable 5.2, Opus 5.5, and Sonnet 5.5 (with a prompt-length price step above 100K tokens). Only the legacy **Haiku 4.5** remains limited to a 200,000 token context window.

Need this set up for your team?

Talk to Will →

■ Deploying Claude or AI Infrastructure in Your Business?

At Tygart Media, we engineer custom Model Context Protocol (MCP) servers, multi-model content pipelines, and AI operational systems. Explore our Claude AI Team Implementation Services or check out our complete Restoration Operations & AI Kit.

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: deprecation guide
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-tokens-as-building-blocks.jpg
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-context-vs-output-bars.jpg
- Tygart Media: Claude Rate Limits: What They Are and How to Work Around Them
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-context-limit-soft-edge.jpg
- Tygart Media: prompt caching
- Tygart Media: Talk to Will →
- Tygart Media: Restoration Operations & AI Kit