Pricing

Learn about Anthropic's pricing structure for models and features

Copy page

This page provides detailed pricing information for Anthropic's models and features. All prices are in USD.

For the most current pricing information, please visit claude.com/pricing.

Model pricing

The following table shows pricing for all Claude models across different usage tiers:

Model	Base Input Tokens	5m Cache Writes	1h Cache Writes	Cache Hits & Refreshes	Output Tokens
Claude Opus 4.5	$5 / MTok	$6.25 / MTok	$10 / MTok	$0.50 / MTok	$25 / MTok
Claude Opus 4.1	$15 / MTok	$18.75 / MTok	$30 / MTok	$1.50 / MTok	$75 / MTok
Claude Opus 4	$15 / MTok	$18.75 / MTok	$30 / MTok	$1.50 / MTok	$75 / MTok
Claude Sonnet 4.5	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok
Claude Sonnet 4	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok
Claude Sonnet 3.7 (deprecated)	$3 / MTok	$3.75 / MTok	$6 / MTok	$0.30 / MTok	$15 / MTok
Claude Haiku 4.5	$1 / MTok	$1.25 / MTok	$2 / MTok	$0.10 / MTok	$5 / MTok
Claude Haiku 3.5	$0.80 / MTok	$1 / MTok	$1.6 / MTok	$0.08 / MTok	$4 / MTok
Claude Opus 3 (deprecated)	$15 / MTok	$18.75 / MTok	$30 / MTok	$1.50 / MTok	$75 / MTok
Claude Haiku 3	$0.25 / MTok	$0.30 / MTok	$0.50 / MTok	$0.03 / MT	

pricing. Prompt caching offers both 5-minute (default) and 1-hour cache durations to optimize costs for different use cases.

The table above reflects the following pricing multipliers for prompt caching:

- 5-minute cache write tokens are 1.25 times the base input tokens price
- 1-hour cache write tokens are 2 times the base input tokens price
- Cache read tokens are 0.1 times the base input tokens price

Third-party platform pricing

Claude models are available on AWS Bedrock, Google Vertex AI, and Microsoft Foundry. For official pricing, visit:

- AWS Bedrock pricing
- Google Vertex AI pricing
- Microsoft Foundry pricing

Regional endpoint pricing for Claude 4.5 models and beyond

Starting with Claude Sonnet 4.5 and Haiku 4.5, AWS Bedrock and Google Vertex AI offer two endpoint types:

- **Global endpoints**: Dynamic routing across regions for maximum availability
- **Regional endpoints**: Data routing guaranteed within specific geographic regions

Regional endpoints include a 10% premium over global endpoints. **The Claude API (1P) is global by default and unaffected by this change.** The Claude API is global-only (equivalent to the global endpoint offering and pricing from other providers).

Scope: This pricing structure applies to Claude Sonnet 4.5, Haiku 4.5, and all future models. Earlier models (Claude Sonnet 4, Opus 4, and prior releases) retain their existing pricing.

For implementation details and code examples:

- AWS Bedrock global vs regional endpoints
- Google Vertex AI global vs regional endpoints

Feature-specific pricing

Claude Docs

The Batch API allows asynchronous processing of large volumes of requests with a 50% discount on both input and output tokens.

Model	Batch input	Batch output
Claude Opus 4.5	$2.50 / MTok	$12.50 / MTok
Claude Opus 4.1	$7.50 / MTok	$37.50 / MTok
Claude Opus 4	$7.50 / MTok	$37.50 / MTok
Claude Sonnet 4.5	$1.50 / MTok	$7.50 / MTok
Claude Sonnet 4	$1.50 / MTok	$7.50 / MTok
Claude Sonnet 3.7 (deprecated)	$1.50 / MTok	$7.50 / MTok
Claude Haiku 4.5	$0.50 / MTok	$2.50 / MTok
Claude Haiku 3.5	$0.40 / MTok	$2 / MTok
Claude Opus 3 (deprecated)	$7.50 / MTok	$37.50 / MTok
Claude Haiku 3	$0.125 / MTok	$0.625 / MTok

For more information about batch processing, see our batch processing documentation.

Long context pricing

When using Claude Sonnet 4 or Sonnet 4.5 with the 1M token context window enabled, requests that exceed 200K input tokens are automatically charged at premium long context rates:

> The 1M token context window is currently in beta for organizations in usage tier 4 and organizations with custom rate limits. The 1M token context window is only available for Claude Sonnet 4 and Sonnet 4.5.

≤ 200K input tokens	> 200K input tokens
Input: $3 / MTok	Input: $6 / MTok
Output: $15 / MTok	Output: $22.50 / MTok

Long context pricing stacks with other pricing modifiers:

> Even with the beta flag enabled, requests with fewer than 200K input tokens are charged at standard rates. If your request exceeds 200K input tokens, all tokens incur premium pricing.
>
> The 200K threshold is based solely on input tokens (including cache reads/writes). Output token count does not affect pricing tier selection, though output tokens are charged at the higher rate when the input threshold is exceeded.

To check if your API request was charged at the 1M context window rates, examine the `usage` object in the API response:

```
{
  "usage": {
    "input_tokens": 250000,
    "cache_creation_input_tokens": 0,
    "cache_read_input_tokens": 0,
    "output_tokens": 500
  }
}
```

Calculate the total input tokens by summing:

- `input_tokens`
- `cache_creation_input_tokens` (if using prompt caching)
- `cache_read_input_tokens` (if using prompt caching)

If the total exceeds 200,000 tokens, the entire request was billed at 1M context rates.

For more information about the `usage` object, see the API response documentation.

Tool use pricing

Tool use requests are priced based on:

1. The total number of input tokens sent to the model (including in the `tools` parameter)
2. The number of output tokens generated
3. For server-side tools, additional usage-based pricing (e.g., web search charges per search performed)

The additional tokens from tool use come from:

- The `tools` parameter in API requests (tool names, descriptions, and schemas)
- `tool_use` content blocks in API requests and responses
- `tool_result` content blocks in API requests

When you use `tools`, we also automatically include a special system prompt for the model which enables tool use. The number of tool use tokens required for each model are listed below (excluding the additional tokens listed above). Note that the table assumes at least 1 tool is provided. If no `tools` are provided, then a tool choice of `none` uses 0 additional system prompt tokens.

Model	Tool choice	Tool use system prompt token count
Claude Opus 4.5	`auto`, `none`	346 tokens
	`any`, `tool`	313 tokens
Claude Opus 4.1	`auto`, `none`	346 tokens
	`any`, `tool`	313 tokens
Claude Opus 4	`auto`, `none`	346 tokens
	`any`, `tool`	313 tokens
Claude Sonnet 4.5	`auto`, `none`	346 tokens
	`any`, `tool`	313 tokens
Claude Sonnet 4	`auto`, `none`	346 tokens
	`any`, `tool`	313 tokens
Claude Sonnet 3.7 (deprecated)	`auto`, `none`	346 tokens
	`any`, `tool`	313 tokens
Claude Haiku 4.5	`auto`, `none`	346 tokens
	`any`, `tool`	313 tokens
Claude Haiku 3.5	`auto`, `none`	264 tokens
	`any`, `tool`	340 tokens
Claude Opus 3 (deprecated)	`auto`, `none`	530 tokens

Claude Sonnet 3	`auto , none`	159 tokens
	`any , tool`	235 tokens
Claude Haiku 3	`auto , none`	264 tokens
	`any , tool`	340 tokens

These token counts are added to your normal input and output tokens to calculate the total cost of a request.

For current per-model prices, refer to our model pricing section above.

For more information about tool use implementation and best practices, see our tool use documentation.

Specific tool pricing

Bash tool

The bash tool adds **245 input tokens** to your API calls.

Additional tokens are consumed by:

- Command outputs (stdout/stderr)
- Error messages
- Large file contents

See tool use pricing for complete pricing details.

Code execution tool

Code execution tool usage is tracked separately from token usage. Execution time has a minimum of 5 minutes. If files are included in the request, execution time is billed even if the tool is not used due to files being preloaded onto the container.

Each organization receives 1,550 free hours of usage with the code execution tool per month. Additional usage beyond the first 1,550 hours is billed at $0.05 per hour, per container.

Text editor tool

using.

In addition to the base tokens, the following additional input tokens are needed for the text editor tool:

Tool	Additional input tokens
`text_editor_20250429` (Claude 4.x)	700 tokens
`text_editor_20250124` (Claude Sonnet 3.7 (deprecated))	700 tokens

See tool use pricing for complete pricing details.

Web search tool

Web search usage is charged in addition to token usage:

```
"usage": {
  "input_tokens": 105,
  "output_tokens": 6039,
  "cache_read_input_tokens": 7123,
  "cache_creation_input_tokens": 7345,
  "server_tool_use": {
    "web_search_requests": 1
  }
}
```

Web search is available on the Claude API for **$10 per 1,000 searches**, plus standard token costs for search-generated content. Web search results retrieved throughout a conversation are counted as input tokens, in search iterations executed during a single turn and in subsequent conversation turns.

Each web search counts as one use, regardless of the number of results returned. If an error occurs during web search, the web search will not be billed.

Web fetch tool

Web fetch usage has **no additional charges** beyond standard token costs:

```
    "output_tokens": 931,
    "cache_read_input_tokens": 0,
    "cache_creation_input_tokens": 0,
    "server_tool_use": {
      "web_fetch_requests": 1
    }
  }
```

The web fetch tool is available on the Claude API at **no additional cost**. You only pay standard token costs for the fetched content that becomes part of your conversation context.

To protect against inadvertently fetching large content that would consume excessive tokens, use the `max_content_tokens` parameter to set appropriate limits based on your use case and budget considerations.

Example token usage for typical content:

- Average web page (10KB): ~2,500 tokens
- Large documentation page (100KB): ~25,000 tokens
- Research paper PDF (500KB): ~125,000 tokens

Computer use tool

Computer use follows the standard tool use pricing. When using the computer use tool:

System prompt overhead: The computer use beta adds 466-499 tokens to the system prompt

Computer use tool token usage:

Model	Input tokens per tool definition
Claude 4.x models	735 tokens
Claude Sonnet 3.7 (deprecated)	735 tokens

Additional token consumption:

- Screenshot images (see Vision pricing)
- Tool execution results returned to Claude

Claude Docs

Agent use case pricing examples

Understanding pricing for agent applications is crucial when building with Claude. These real-world examples can help you estimate costs for different agent patterns.

Customer support agent example

When building a customer support agent, here's how costs might break down:

Example calculation for processing 10,000 support tickets:

- Average ~3,700 tokens per conversation
- Using Claude Sonnet 4.5 at $3/MTok input, $15/MTok output
- Total cost: ~$22.20 per 10,000 tickets

For a detailed walkthrough of this calculation, see our customer support agent guide.

General agent workflow pricing

For more complex agent architectures with multiple steps:

1. **Initial request processing**
   - Typical input: 500-1,000 tokens
   - Processing cost: ~$0.003 per request
2. **Memory and context retrieval**
   - Retrieved context: 2,000-5,000 tokens
   - Cost per retrieval: ~$0.015 per operation
3. **Action planning and execution**
   - Planning tokens: 1,000-2,000
   - Execution feedback: 500-1,000
   - Combined cost: ~$0.045 per action

For a comprehensive guide on agent pricing patterns, see our agent use cases guide.

When building agents with Claude:

1. **Use appropriate models**: Choose Haiku for simple tasks, Sonnet for complex reasoning
2. **Implement prompt caching**: Reduce costs for repeated context
3. **Batch operations**: Use the Batch API for non-time-sensitive tasks
4. **Monitor usage patterns**: Track token consumption to identify optimization opportunities

> For high-volume agent applications, consider contacting our enterprise sales team for custom pricing arrangements.

Additional pricing considerations

Rate limits

Rate limits vary by usage tier and affect how many requests you can make:

- **Tier 1**: Entry-level usage with basic limits
- **Tier 2**: Increased limits for growing applications
- **Tier 3**: Higher limits for established applications
- **Tier 4**: Maximum standard limits
- **Enterprise**: Custom limits available

For detailed rate limit information, see our rate limits documentation.

For higher rate limits or custom pricing arrangements, contact our sales team.

Volume discounts

Volume discounts may be available for high-volume users. These are negotiated on a case-by-case basis.

- Standard tiers use the pricing shown above
- Enterprise customers can contact sales for custom pricing
- Academic and research discounts may be available

For enterprise customers with specific needs:

- Custom rate limits
- Volume discounts
- Dedicated support
- Custom terms

Contact our sales team at sales@anthropic.com or through the Claude Console to discuss enterprise pricing options.

Billing and payment

- Billing is calculated monthly based on actual usage
- Payments are processed in USD
- Credit card and invoicing options available
- Usage tracking available in the Claude Console

Frequently asked questions

How is token usage calculated?

Tokens are pieces of text that models process. As a rough estimate, 1 token is approximately 4 characters or 0.75 words in English. The exact count varies by language and content type.

Are there free tiers or trials?

New users receive a small amount of free credits to test the API. Contact sales for information about extended trials for enterprise evaluation.

How do discounts stack?

Batch API and prompt caching discounts can be combined. For example, using both features together provides significant cost savings compared to standard API calls.

What payment methods are accepted?

We accept major credit cards for standard accounts. Enterprise customers can arrange invoicing and other payment methods.

Claude Docs

Claude Docs

Solutions

AI agents

Code modernization

Coding

Customer support

Education

Financial services

Government

Life sciences

Partners

Amazon Bedrock

Google Cloud's Vertex AI

Learn

Blog

Catalog

Courses

Use cases

Connectors

Customer stories

Engineering at Anthropic

Events

Powered by Claude

Service partners

Startups program

Company

Anthropic

Careers

Economic Futures

Research

News

Responsible Scaling Policy

Security and compliance

Transparency

Help and security

Availability

Status

Support

Discord

Terms and policies

Claude Docs

Terms of service: Commercial

Terms of service: Consumer

Usage policy