Claude Model Pricing — List Prices by Platform

Prices Effective: October 22, 2026 (2026-10-22)

This document contains pricing information as of the effective date listed above. All prices are USD per one million tokens (MTok) unless otherwise noted. "—" indicates the SKU is not offered for that pricing tier / scope / context window combination. Rows marked "*" indicate that SKU availability varies by region within the listed scope; not all SKUs in that row are offered in every region.

This rate card sets forth certain list prices. Anthropic may describe promotional pricing elsewhere, including on anthropic.com and claude.com. Promotional pricing is not a list price, does not modify the list prices set forth in this rate card, and applies only for the period and on the terms of the applicable promotion.

For further information on all other feature-, tool- and model-specific pricing visit: https://platform.claude.com/docs/en/about-claude/pricing

Claude API

15 model SKUs · prices in USD per 1M tokens

Claude Opus 5.5 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$4.00	$20.00	$5.00	$8.00	$0.20
Standard	US-only inference	All	$4.40	$22.00	$5.50	$8.80	$0.22
Batch processing	Global	All	$2.00	$10.00	$2.50	$4.00	$0.10
Batch processing	US-only inference	All	$2.20	$11.00	$2.75	$4.40	$0.11
Fast mode	Global	All	$8.00	$40.00	$10.00	$16.00	$0.40
Fast mode	US-only inference	All	$8.80	$44.00	$11.00	$17.60	$0.44

Claude Sonnet 5.5 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$2.00	$10.00	$2.50	$4.00	$0.10
Standard	US-only inference	All	$2.20	$11.00	$2.75	$4.40	$0.11
Batch processing	Global	All	$1.00	$5.00	$1.25	$2.00	$0.05
Batch processing	US-only inference	All	$1.10	$5.50	$1.375	$2.20	$0.055

Claude Haiku 5.5 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	≤100K	$0.10	$0.50	$0.125	$0.20	$0.01
Standard	Global	>100K	$0.50	$2.50	$0.625	$1.00	$0.05
Standard	US-only inference	≤100K	$0.11	$0.55	$0.1375	$0.22	$0.011
Standard	US-only inference	>100K	$0.55	$2.75	$0.6875	$1.10	$0.055

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Batch processing	Global	≤100K	$0.05	$0.25	$0.0625	$0.10	$0.005
Batch processing	Global	>100K	$0.25	$1.25	$0.3125	$0.50	$0.025
Batch processing	US-only inference	≤100K	$0.055	$0.275	$0.06875	$0.11	$0.0055
Batch processing	US-only inference	>100K	$0.275	$1.375	$0.34375	$0.55	$0.0275

Claude Mythos 5.1 (limited availability) — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$0.25
Standard	US-only inference	All	$11.00	$55.00	$13.75	$22.00	$0.275
Batch processing	Global	All	$5.00	$25.00	$6.25	$10.00	$0.125
Batch processing	US-only inference	All	$5.50	$27.50	$6.875	$11.00	$0.1375

Claude Fable 5.1 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$0.25
Standard	US-only inference	All	$11.00	$55.00	$13.75	$22.00	$0.275
Batch processing	Global	All	$5.00	$25.00	$6.25	$10.00	$0.125
Batch processing	US-only inference	All	$5.50	$27.50	$6.875	$11.00	$0.1375

Claude Opus 5 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	US-only inference	All	$5.50	$27.50	$6.875	$11.00	$0.55
Batch processing	Global	All	$2.50	$12.50	$3.125	$5.00	$0.25
Batch processing	US-only inference	All	$2.75	$13.75	$3.4375	$5.50	$0.275

Claude Sonnet 5 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$2.00	$10.00	$2.50	$4.00	$0.20
Standard	US-only inference	All	$2.20	$11.00	$2.75	$4.40	$0.22

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Batch processing	Global	All	$1.00	$5.00	$1.25	$2.00	$0.10
Batch processing	US-only inference	All	$1.10	$5.50	$1.375	$2.20	$0.11

Claude Mythos 5 (limited availability) — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$1.00
Standard	US-only inference	All	$11.00	$55.00	$13.75	$22.00	$1.10
Batch processing	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Batch processing	US-only inference	All	$5.50	$27.50	$6.875	$11.00	$0.55

Claude Fable 5 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$1.00
Standard	US-only inference	All	$11.00	$55.00	$13.75	$22.00	$1.10
Batch processing	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Batch processing	US-only inference	All	$5.50	$27.50	$6.875	$11.00	$0.55

Claude Opus 4.8 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	US-only inference	All	$5.50	$27.50	$6.875	$11.00	$0.55
Batch processing	Global	All	$2.50	$12.50	$3.125	$5.00	$0.25
Batch processing	US-only inference	All	$2.75	$13.75	$3.4375	$5.50	$0.275

Claude Opus 4.7 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	US-only inference	All	$5.50	$27.50	$6.875	$11.00	$0.55
Batch processing	Global	All	$2.50	$12.50	$3.125	$5.00	$0.25
Batch processing	US-only inference	All	$2.75	$13.75	$3.4375	$5.50	$0.275

Claude Opus 4.6 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	US-only inference	All	$5.50	$27.50	$6.875	$11.00	$0.55
Batch processing	Global	All	$2.50	$12.50	$3.125	$5.00	$0.25
Batch processing	US-only inference	All	$2.75	$13.75	$3.4375	$5.50	$0.275

Claude Sonnet 4.6 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$3.00	$15.00	$3.75	$6.00	$0.30
Standard	US-only inference	All	$3.30	$16.50	$4.125	$6.60	$0.33
Batch processing	Global	All	$1.50	$7.50	$1.875	$3.00	$0.15
Batch processing	US-only inference	All	$1.65	$8.25	$2.0625	$3.30	$0.165

Claude Opus 4.5 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	≤200K	$5.00	$25.00	$6.25	$10.00	$0.50
Batch processing	Global	≤200K	$2.50	$12.50	$3.125	$5.00	$0.25

Claude Haiku 4.5 — Claude API

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	≤200K	$1.00	$5.00	$1.25	$2.00	$0.10
Batch processing	Global	≤200K	$0.50	$2.50	$0.625	$1.00	$0.05

> Claude Managed Agents: Customer's use of Claude Managed Agents is billed at the standard model token rates above plus a session-runtime fee of $0.08 USD per session-hour, metered while a session's status is "running".

> If Customer accesses Claude API through Claude Platform on AWS, Claude in Microsoft Foundry, or Claude Models on Google Cloud (Anthropic Managed) (each, a Marketplace Platform), usage will be invoiced in Claude Consumption Units ("CCU"). A CCU is a unit of measure used for the Marketplace Platform's invoicing. One hundred (100) CCU represents $1.00 USD of fees owed for the Services.

AWS Bedrock

15 model SKUs · prices in USD per 1M tokens

Claude Opus 5.5 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$4.00	$20.00	$5.00	$8.00	$0.20
Standard	In-Region & Geo Cross Region	All	$4.40	$22.00	$5.50	$8.80	$0.22
Standard	AWS GovCloud (US)	All	$4.80	$24.00	$6.00	$9.60	$0.24

Claude Sonnet 5.5 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$2.00	$10.00	$2.50	$4.00	$0.10
Standard	In-Region & Geo Cross Region	All	$2.20	$11.00	$2.75	$4.40	$0.11
Standard	AWS GovCloud (US)	All	$2.40	$12.00	$3.00	$4.80	$0.12

Claude Haiku 5.5 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	≤100K	$0.10	$0.50	$0.125	$0.20	$0.01
Standard	Global Cross Region	>100K	$0.50	$2.50	$0.625	$1.00	$0.05
Standard	In-Region & Geo Cross Region	≤100K	$0.11	$0.55	$0.1375	$0.22	$0.011
Standard	In-Region & Geo Cross Region	>100K	$0.55	$2.75	$0.6875	$1.10	$0.055
Standard	AWS GovCloud (US)	≤100K	$0.12	$0.60	$0.15	$0.24	$0.012
Standard	AWS GovCloud (US)	>100K	$0.60	$3.00	$0.75	$1.20	$0.06

Claude Mythos 5.1 (limited availability) — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$10.00	$50.00	$12.50	$20.00	$0.25
Standard	In-Region & Geo Cross Region	All	$11.00	$55.00	$13.75	$22.00	$0.275

Claude Fable 5.1 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$10.00	$50.00	$12.50	$20.00	$0.25
Standard	In-Region & Geo Cross Region	All	$11.00	$55.00	$13.75	$22.00	$0.275
Standard	AWS GovCloud (US)	All	$12.00	$60.00	$15.00	$24.00	$0.30

Claude Opus 5 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	In-Region & Geo Cross Region	All	$5.50	$27.50	$6.875	$11.00	$0.55
Standard	AWS GovCloud (US)	All	$6.00	$30.00	$7.50	$12.00	$0.60
Batch processing	Global Cross Region	All	$2.50	$12.50	—	—	—
Batch processing	In-Region & Geo Cross Region	All	$2.75	$13.75	—	—	—
Batch processing	AWS GovCloud (US)	All	$3.00	$15.00	—	—	—

Claude Sonnet 5 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$2.00	$10.00	$2.50	$4.00	$0.20
Standard	In-Region & Geo Cross Region	All	$2.20	$11.00	$2.75	$4.40	$0.22
Standard	AWS GovCloud (US)	All	$2.40	$12.00	$3.00	$4.80	$0.24
Batch processing	AWS GovCloud (US)	All	$1.20	$6.00	—	—	—

Claude Mythos 5 (limited availability) — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	In-Region & Geo Cross Region	All	$11.00	$55.00	$13.75	$22.00	$1.10

Claude Fable 5 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$10.00	$50.00	$12.50	$20.00	$1.00
Standard	In-Region & Geo Cross Region	All	$11.00	$55.00	$13.75	$22.00	$1.10

Claude Opus 4.8 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	In-Region & Geo Cross Region	All	$5.50	$27.50	$6.875	$11.00	$0.55
Standard	AWS GovCloud (US)	All	$6.00	$30.00	$7.50	$12.00	$0.60
Batch processing	Global Cross Region	All	$2.50	$12.50	—	—	—

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Batch processing	In-Region & Geo Cross Region	All	$2.75	$13.75	—	—	—

Claude Opus 4.7 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	In-Region & Geo Cross Region	All	$5.50	$27.50	$6.875	$11.00	$0.55
Batch processing	Global Cross Region	All	$2.50	$12.50	—	—	—
Batch processing	In-Region & Geo Cross Region	All	$2.75	$13.75	—	—	—

Claude Opus 4.6 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	In-Region & Geo Cross Region	All	$5.50	$27.50	$6.875	$11.00	$0.55
Standard	AWS GovCloud (US)	All	$6.00	$30.00	$7.50	—	$0.60
Batch processing	Global Cross Region	All	$2.50	$12.50	—	—	—
Batch processing	In-Region & Geo Cross Region	All	$2.75	$13.75	—	—	—
Batch processing	AWS GovCloud (US)	All	$3.00	$15.00	—	—	—

Claude Sonnet 4.6 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	All	$3.00	$15.00	$3.75	$6.00	$0.30
Standard	In-Region & Geo Cross Region	All	$3.30	$16.50	$4.125	$6.60	$0.33
Standard	AWS GovCloud (US)	All	$3.60	$18.00	$4.50	—	$0.36
Batch processing	Global Cross Region	All	$1.50	$7.50	—	—	—
Batch processing	In-Region & Geo Cross Region	All	$1.65	$8.25	—	—	—
Batch processing	AWS GovCloud (US)	All	$1.80	$9.00	—	—	—

Claude Opus 4.5 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	≤200K	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	In-Region & Geo Cross Region *	≤200K	$5.50	$27.50	$6.875	$11.00	$0.55
Standard	In-Region & Geo Cross Region *	≤200K	$5.50	$27.50	$6.875	—	$0.55
Batch processing	Global Cross Region	≤200K	$2.50	$12.50	—	—	—
Batch processing	In-Region & Geo Cross Region	≤200K	$2.75	$13.75	—	—	—

Claude Haiku 4.5 — AWS Bedrock

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global Cross Region	≤200K	$1.00	$5.00	$1.25	$2.00	$0.10
Standard	In-Region & Geo Cross Region	≤200K	$1.10	$5.50	$1.375	$2.20	$0.11
Batch processing	Global Cross Region	≤200K	$0.50	$2.50	—	—	—
Batch processing	In-Region & Geo Cross Region	≤200K	$0.55	$2.75	—	—	—

Google Vertex AI

15 model SKUs · prices in USD per 1M tokens

Claude Opus 5.5 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$4.00	$20.00	$5.00	$8.00	$0.20
Standard	Multi-Region and Regional Endpoint	All	$4.40	$22.00	$5.50	$8.80	$0.22
Batch processing	Global	All	$2.00	$10.00	$2.50	$4.00	$0.10
Batch processing	Multi-Region and Regional Endpoint	All	$2.20	$11.00	$2.75	$4.40	$0.11

Claude Sonnet 5.5 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$2.00	$10.00	$2.50	$4.00	$0.10
Standard	Multi-Region and Regional Endpoint	All	$2.20	$11.00	$2.75	$4.40	$0.11
Batch processing	Global	All	$1.00	$5.00	$1.25	$2.00	$0.05
Batch processing	Multi-Region and Regional Endpoint	All	$1.10	$5.50	$1.375	$2.20	$0.055

Claude Haiku 5.5 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	≤100K	$0.10	$0.50	$0.125	$0.20	$0.01
Standard	Global	>100K	$0.50	$2.50	$0.625	$1.00	$0.05
Standard	Multi-Region and Regional Endpoint	≤100K	$0.11	$0.55	$0.1375	$0.22	$0.011
Standard	Multi-Region and Regional Endpoint	>100K	$0.55	$2.75	$0.6875	$1.10	$0.055
Batch processing	Global	≤100K	$0.05	$0.25	$0.0625	$0.10	$0.005
Batch processing	Global	>100K	$0.25	$1.25	$0.3125	$0.50	$0.025
Batch processing	Multi-Region and Regional Endpoint	≤100K	$0.055	$0.275	$0.06875	$0.11	$0.0055
Batch processing	Multi-Region and Regional Endpoint	>100K	$0.275	$1.375	$0.34375	$0.55	$0.0275

Claude Mythos 5.1 (limited availability) — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$0.25

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Multi-Region and Regional Endpoint	All	$11.00	$55.00	$13.75	$22.00	$0.275
Batch processing	Global	All	$5.00	$25.00	$6.25	$10.00	$0.125
Batch processing	Multi-Region and Regional Endpoint	All	$5.50	$27.50	$6.875	$11.00	$0.1375

Claude Fable 5.1 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$0.25
Standard	Multi-Region and Regional Endpoint	All	$11.00	$55.00	$13.75	$22.00	$0.275
Batch processing	Global	All	$5.00	$25.00	$6.25	$10.00	$0.125
Batch processing	Multi-Region and Regional Endpoint	All	$5.50	$27.50	$6.875	$11.00	$0.1375

Claude Opus 5 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	Multi-Region and Regional Endpoint	All	$5.50	$27.50	$6.875	$11.00	$0.55
Batch processing	Global	All	$2.50	$12.50	$3.125	$5.00	$0.25
Batch processing	Multi-Region and Regional Endpoint	All	$2.75	$13.75	$3.4375	$5.50	$0.275

Claude Sonnet 5 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$2.00	$10.00	$2.50	$4.00	$0.20
Standard	Multi-Region and Regional Endpoint	All	$2.20	$11.00	$2.75	$4.40	$0.22
Batch processing	Global	All	$1.00	$5.00	$1.25	$2.00	$0.10
Batch processing	Multi-Region and Regional Endpoint	All	$1.10	$5.50	$1.375	$2.20	$0.11

Claude Mythos 5 (limited availability) — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$1.00

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Multi-Region and Regional Endpoint	All	$11.00	$55.00	$13.75	$22.00	$1.10
Batch processing	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Batch processing	Multi-Region and Regional Endpoint	All	$5.50	$27.50	$6.875	$11.00	$0.55

Claude Fable 5 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$1.00
Standard	Multi-Region and Regional Endpoint	All	$11.00	$55.00	$13.75	$22.00	$1.10
Batch processing	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Batch processing	Multi-Region and Regional Endpoint	All	$5.50	$27.50	$6.875	$11.00	$0.55

Claude Opus 4.8 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	Multi-Region and Regional Endpoint	All	$5.50	$27.50	$6.875	$11.00	$0.55
Batch processing	Global	All	$2.50	$12.50	$3.125	$5.00	$0.25
Batch processing	Multi-Region and Regional Endpoint	All	$2.75	$13.75	$3.4375	$5.50	$0.275

Claude Opus 4.7 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	Multi-Region and Regional Endpoint	All	$5.50	$27.50	$6.875	$11.00	$0.55
Batch processing	Global	All	$2.50	$12.50	$3.125	$5.00	$0.25
Batch processing	Multi-Region and Regional Endpoint	All	$2.75	$13.75	$3.4375	$5.50	$0.275

Claude Opus 4.6 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Multi-Region and Regional Endpoint	All	$5.50	$27.50	$6.875	$11.00	$0.55
Batch processing	Global	All	$2.50	$12.50	$3.125	$5.00	$0.25
Batch processing	Multi-Region and Regional Endpoint	All	$2.75	$13.75	$3.4375	$5.50	$0.275

Claude Sonnet 4.6 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$3.00	$15.00	$3.75	$6.00	$0.30
Standard	Multi-Region and Regional Endpoint	All	$3.30	$16.50	$4.125	$6.60	$0.33
Batch processing	Global	All	$1.50	$7.50	$1.875	$3.00	$0.15
Batch processing	Multi-Region and Regional Endpoint	All	$1.65	$8.25	$2.0625	$3.30	$0.165

Claude Opus 4.5 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	≤200K	$5.00	$25.00	$6.25	$10.00	$0.50
Standard	Multi-Region and Regional Endpoint	≤200K	$5.50	$27.50	$6.875	$11.00	$0.55
Batch processing	Global	≤200K	$2.50	$12.50	$3.125	$5.00	$0.25
Batch processing	Multi-Region and Regional Endpoint	≤200K	$2.75	$13.75	$3.4375	$5.50	$0.275

Claude Haiku 4.5 — Google Vertex AI

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	≤200K	$1.00	$5.00	$1.25	$2.00	$0.10
Standard	Multi-Region and Regional Endpoint	≤200K	$1.10	$5.50	$1.375	$2.20	$0.11
Batch processing	Global	≤200K	$0.50	$2.50	$0.625	$1.00	$0.05
Batch processing	Multi-Region and Regional Endpoint	≤200K	$0.55	$2.75	$0.6875	$1.10	$0.055

Microsoft Foundry

15 model SKUs · prices in USD per 1M tokens

Claude Opus 5.5 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$4.00	$20.00	$5.00	$8.00	$0.20

Claude Sonnet 5.5 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$2.00	$10.00	$2.50	$4.00	$0.10

Claude Haiku 5.5 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	≤100K	$0.10	$0.50	$0.125	$0.20	$0.01
Standard	Global	>100K	$0.50	$2.50	$0.625	$1.00	$0.05

Claude Mythos 5.1 (limited availability) — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$0.25

Claude Fable 5.1 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$0.25

Claude Opus 5 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50

Claude Sonnet 5 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$2.00	$10.00	$2.50	$4.00	$0.20

Claude Mythos 5 (limited availability) — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$1.00

Claude Fable 5 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$10.00	$50.00	$12.50	$20.00	$1.00

Claude Opus 4.8 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50

Claude Opus 4.7 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50

Claude Opus 4.6 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$5.00	$25.00	$6.25	$10.00	$0.50

Claude Sonnet 4.6 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	All	$3.00	$15.00	$3.75	$6.00	$0.30

Claude Opus 4.5 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	≤200K	$5.00	$25.00	$6.25	$10.00	$0.50

Claude Haiku 4.5 — Microsoft Foundry

PRICING TIER	SCOPE	CONTEXT WINDOW	BASE INPUT TOKENS	OUTPUT TOKENS	5M CACHE WRITES	1H CACHE WRITES	CACHE HITS & REFRESHES
Standard	Global	≤200K	$1.00	$5.00	$1.25	$2.00	$0.10