

# Claude Model Pricing — List Prices by Platform

Prices Effective: June 30, 2026 (2026-06-30)

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](https://anthropic.com) and [claude.com](https://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

9 model SKUs · prices in USD per 1M tokens

### 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            | \$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 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 |
|------------------|-------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global            | All            | \$5.00            | \$25.00       | \$6.25          | \$10.00         | \$0.50                 |
| <b>Standard</b>  | 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 |
|------------------|-------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global            | All            | \$5.00            | \$25.00       | \$6.25          | \$10.00         | \$0.50                 |
| <b>Standard</b>  | 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 |
|------------------|-------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global            | All            | \$3.00            | \$15.00       | \$3.75          | \$6.00          | \$0.30                 |
| <b>Standard</b>  | 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 |
|------------------|--------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | 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 Sonnet 4.5 — Claude API

| PRICING TIER     | SCOPE  | CONTEXT WINDOW | BASE INPUT TOKENS | OUTPUT TOKENS | 5M CACHE WRITES | 1H CACHE WRITES | CACHE HITS & REFRESHES |
|------------------|--------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global | ≤200K          | \$3.00            | \$15.00       | \$3.75          | \$6.00          | \$0.30                 |
| Batch processing | Global | ≤200K          | \$1.50            | \$7.50        | \$1.875         | \$3.00          | \$0.15                 |

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 or Claude in Microsoft Foundry (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

9 model SKUs · prices in USD per 1M tokens

### 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            | \$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                 |

### 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                 |

### 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                 |

### 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 |
|------------------|------------------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global Cross Region          | All            | \$3.00            | \$15.00       | \$3.75          | \$6.00          | \$0.30                 |
| <b>Standard</b>  | In-Region & Geo Cross Region | All            | \$3.30            | \$16.50       | \$4.125         | \$6.60          | \$0.33                 |
| <b>Standard</b>  | 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 |
|------------------|--------------------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global Cross Region            | ≤200K          | \$5.00            | \$25.00       | \$6.25          | \$10.00         | \$0.50                 |
| <b>Standard</b>  | In-Region & Geo Cross Region * | ≤200K          | \$5.50            | \$27.50       | \$6.875         | \$11.00         | \$0.55                 |
| <b>Standard</b>  | 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 Sonnet 4.5 — AWS Bedrock

| PRICING TIER     | SCOPE                        | CONTEXT WINDOW | BASE INPUT TOKENS | OUTPUT TOKENS | 5M CACHE WRITES | 1H CACHE WRITES | CACHE HITS & REFRESHES |
|------------------|------------------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global Cross Region          | ≤200K          | \$3.00            | \$15.00       | \$3.75          | \$6.00          | \$0.30                 |
| <b>Standard</b>  | In-Region & Geo Cross Region | ≤200K          | \$3.30            | \$16.50       | \$4.125         | \$6.60          | \$0.33                 |
| <b>Standard</b>  | AWS GovCloud (US)            | ≤200K          | \$3.60            | \$18.00       | \$4.50          | \$7.20          | \$0.36                 |
| Batch processing | Global Cross Region          | ≤200K          | \$1.50            | \$7.50        | —               | —               | —                      |
| Batch processing | In-Region & Geo Cross Region | ≤200K          | \$1.65            | \$8.25        | —               | —               | —                      |
| Batch processing | AWS GovCloud (US)            | ≤200K          | \$1.80            | \$9.00        | —               | —               | —                      |

## 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 |
|-----------------|---------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b> | Global Cross Region | ≤200K          | \$1.00            | \$5.00        | \$1.25          | \$2.00          | \$0.10                 |

| 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 | ≤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

9 model SKUs · prices in USD per 1M tokens

## 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            | \$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 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 |
|------------------|------------------------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global                             | All            | \$5.00            | \$25.00       | \$6.25          | \$10.00         | \$0.50                 |
| <b>Standard</b>  | 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 |
|------------------|------------------------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global                             | All            | \$3.00            | \$15.00       | \$3.75          | \$6.00          | \$0.30                 |
| <b>Standard</b>  | 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 |
|------------------|------------------------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global                             | ≤200K          | \$5.00            | \$25.00       | \$6.25          | \$10.00         | \$0.50                 |
| <b>Standard</b>  | 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 Sonnet 4.5 — Google Vertex AI

| PRICING TIER     | SCOPE                              | CONTEXT WINDOW | BASE INPUT TOKENS | OUTPUT TOKENS | 5M CACHE WRITES | 1H CACHE WRITES | CACHE HITS & REFRESHES |
|------------------|------------------------------------|----------------|-------------------|---------------|-----------------|-----------------|------------------------|
| <b>Standard</b>  | Global                             | ≤200K          | \$3.00            | \$15.00       | \$3.75          | —               | \$0.30                 |
| <b>Standard</b>  | Multi-Region and Regional Endpoint | ≤200K          | \$3.30            | \$16.50       | \$4.125         | —               | \$0.33                 |
| Batch processing | Global                             | ≤200K          | \$1.50            | \$7.50        | \$1.875         | —               | \$0.15                 |
| Batch processing | Multi-Region and Regional Endpoint | ≤200K          | \$1.65            | \$8.25        | \$2.0625        | —               | \$0.165                |



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          | —               | \$0.10                 |
| Standard         | Multi-Region and Regional Endpoint | ≤200K          | \$1.10            | \$5.50        | \$1.375         | —               | \$0.11                 |
| Batch processing | Global                             | ≤200K          | \$0.50            | \$2.50        | \$0.625         | —               | \$0.05                 |
| Batch processing | Multi-Region and Regional Endpoint | ≤200K          | \$0.55            | \$2.75        | \$0.6875        | —               | \$0.055                |

## Microsoft Foundry

9 model SKUs · prices in USD per 1M tokens

### 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            | \$3.00            | \$15.00       | \$3.75          | \$6.00          | \$0.30                 |

### 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 Sonnet 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          | \$3.00            | \$15.00       | \$3.75          | \$6.00          | \$0.30                 |

### 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                 |