Claude API Pricing & Token Rates Schedule (2026)

Edition of November 13, 2026 — facts verified November 13, 2026. Current edition always at https://tygartmedia.com/claude-api-pricing-token-costs-rate-limits-2026/.

Direct Answer (November 13, 2026): Official first-party list: **Fable 5.2 $8 / $40 per MTok** (5m/1h cache writes $10/$16, cache hits $0.20, batch $4/$20); **Opus 5.5 $4 / $20** (cache writes $5/$8, cache hits $0.20); **Sonnet 5.5 $2 / $10** (cache writes $2.50/$4, cache hits $0.10); **Haiku 5.5 $0.10 / $0.50** for prompts ≤100K tokens and **$0.50 / $2.50** >100K (cache hits $0.01/$0.05; batch $0.05/$0.25). Legacy tiers still listed for transition include Fable 5.1 ($10/$50, cache read $0.25), Opus 5 / 4.8 ($5/$25), Sonnet 4.6 ($3/$15), and Haiku 4.5 ($1/$5). Batch API is 50% off. Do not average Haiku 5.5's two prompt-length bands.

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

Updated November 2, 2026 — added direct answers on token pricing schedules and input versus output calculations, sourced from Anthropic's published pricing.

Claude's API pricing is token-based: you pay for the tokens you send (input) and the tokens Claude generates (output). Rate limits, service tiers, prompt caching, batch processing, and feature-specific charges all affect your actual bill. Last refreshed: November 13, 2026 (Pacific) against platform.claude.com pricing and Haiku 5.5. Seat prices did not move. This page includes current Fable 5.2 ($8/$40), Opus 5.5 ($4/$20), Sonnet 5.5 ($2/$10), and Haiku 5.5 ($0.10/$0.50 up to 100K tokens) rates.

Per-Token Pricing by Model

Per-token pricing by model — no sticky dollars.

All prices are per million tokens (MTok), verified November 13, 2026. **Claude Fable 5.2** leads at **$8 input / $40 output** (cache reads $0.20). **Claude Opus 5.5** is **$4 / $20** with cache reads at $0.20. **Claude Sonnet 5.5** is **$2 / $10** with cache reads at $0.10. **Claude Haiku 5.5** lists **$0.10 / $0.50** up to 100,000 prompt tokens and **$0.50 / $2.50** above that. Legacy references still listed: Fable 5.1 ($10/$50, cache reads $0.25), Opus 5 / 4.8 ($5/$25), Sonnet 4.6 ($3/$15), and Haiku 4.5 ($1/$5, 200K).

How much does the Claude API cost per million tokens?

Model	Standard Input (per MTok)	Standard Output (per MTok)	Cache Read (per MTok)	Batch Input (per MTok)	Batch Output (per MTok)
Claude Haiku 5.5 (prompts ≤ 100K tokens)	$0.10	$0.50	$0.01	$0.05	$0.25
Claude Haiku 5.5 (prompts > 100K tokens)	$0.50	$2.50	$0.05	$0.25	$1.25
Claude Haiku 4.5	$1.00	$5.00	$0.10	$0.50	$2.50
Claude Sonnet 5.5	$2.00	$10.00	$0.10	$1.00	$5.00
Claude Sonnet 4.6	$3.00	$15.00	$0.30	$1.50	$7.50
Claude Opus 5.5	$4.00	$20.00	$0.20	$2.00	$10.00
Claude Fable 5.2	$8.00	$40.00	$0.20	$4.00	$20.00
Claude Fable 5.1 (legacy)	$10.00	$50.00	$0.25	$5.00	$25.00

Current Claude API token pricing per million tokens across standard, cached, and batch tiers (sourced from Anthropic published pricing).

What's the difference between input and output token pricing on Claude?

Input tokens represent the prompt text, system instructions, tool declarations, and context history transmitted to Claude, while output tokens represent the new text Claude generates in response. Across all flagship tiers, **Anthropic prices output tokens at exactly 5x the cost of input tokens** (for example, Sonnet 5.5 is $2.00 input vs. $10.00 output; Haiku 5.5 is $0.10 input vs. $0.50 output; Opus 5.5 is $4.00 input vs. $20.00 output). Output tokens carry a 5x multiplier because text generation requires sequential autoregressive inference—each token must be predicted and computed one step at a time—which is substantially more compute-intensive than parallelized input processing.

Worked Calculation: Why Output-Heavy Workloads Dominate Spend Consider 10,000 API calls per day on Claude Sonnet 5.5 ($2.00/MTok input, $10.00/MTok output):

• **Workload A (Input-heavy analysis / summarization):** 4,000 input tokens + 200 output tokens per call. – Daily Input: 40 MTok × $2.00 = $80.00 – Daily Output: 2 MTok × $10.00 = $20.00 – Total Spend: **$100.00/day** (Output accounts for only 20% of cost).

• **Workload B (Output-heavy generation / scaffolding):** 500 input tokens + 3,500 output tokens per call. – Daily Input: 5 MTok × $2.00 = $10.00 – Daily Output: 35 MTok × $10.00 = $350.00 – Total Spend: **$360.00/day** (Output accounts for 97.2% of cost).

Even though Workload B processes fewer total tokens per request than Workload A (4,000 vs. 4,200), it costs **3.6x more** because 87.5% of its tokens fall into the 5x output pricing tier.

Prompt Caching Pricing

5-minute cache write is 1.25x input. 1-hour cache write is 2x input. Cache read is 0.05x on Sonnet 5.5 and Opus 5.5, 0.1x on Sonnet 5 and Haiku 4.5. Fable 5.2 uses $10 5-minute write / $16 1-hour write / $0.20 cache read (0.025x on legacy Fable 5.1 is $0.25/MTok). Haiku 5.5 cache reads are $0.01 / $0.05 (under / over 100k). Haiku 5.5 cache writes are $0.125 / $0.625. Opus 5.5 cache writes $5 / 1-hour $8 / reads $0.20. Opus 5 cache writes $6.25 / reads $0.50. Sonnet 5.5 cache writes $2.50 / reads $0.10. Sonnet 5 cache writes $2.50 / reads $0.20. Sonnet 4.6 cache writes $3.75 / reads $0.30. Haiku 4.5 cache writes $1.25 / reads $0.10.

Batch Processing: 50% Off

The Batch API processes requests asynchronously at half the standard rate. Haiku 5.5 batch is $0.05 / $0.25 under 100k tokens and $0.25 / $1.25 over. Sonnet 5.5 and Sonnet 5 batch list is $1 / $5. Sonnet 4.6 batch list is $1.50 / $7.50. Opus 5.5 batch is $2 / $10. Opus 5 batch is $2.50 / $12.50. Fable 5.2 batch is $4 / $20 (legacy Fable 5.1 batch is $5 / $25).

How to Calculate Your Monthly Bill

To estimate your monthly Claude API bill, calculate your expected daily token volume—multiplying daily requests by average input tokens (at the model's input rate per million) plus daily requests by average output tokens (at the model's output rate per million)—multiply that daily sum by 30 days, and deduct applicable discounts for prompt caching (up to 90% off cached inputs) or batch processing (50% off total spend). Here is how that calculation works in practice:

Example at Sonnet 5.5 list ($2 / $10): 2,000 input tokens and 500 output tokens × 10,000 requests/day = 20 MTok input ($40) + 5 MTok output ($50) = $90/day, about $2,700/month before cache or batch. Same volume on Opus 5.5 at $4 / $20 is $180/day before cache. Same volume on Haiku 5.5 at the under-100k band ($0.10 / $0.50) is $4.50/day before cache.

Same volume with caching on Sonnet 5.5: at a 90% cache-hit rate, 18 of those 20 MTok re-read from cache at $0.10/MTok ($1.80) instead of $2/MTok ($36) — input drops from $40 to $5.80/day. The cache line, not the input line, is usually what separates the sticker price from the real bill. That example used $0.20 before October 7.

Service Tiers and Rate Limits

Priority, Standard, and Batch tiers still apply. US-only inference (`inference_geo: "us"`) on Claude 4.6 and later is 1.1x. Fast mode on Opus 5.5 is $8 / $40. Fast mode on Opus 5 / Opus 4.8 is $10 / $50. Haiku 5.5 is the exception on long-context pricing: prompts over 100,000 tokens use the higher band. Other Claude 4.6 and later models keep one per-token rate through the 1M window. Check live limits in the Claude Console; published RPM/ITPM figures move by spend tier. Confirm five-hour windows in Settings > Usage, not from a screenshot.

Current subscription monthly API credits: Max ($100/mo) includes $100/mo in API credits, Ultra ($200/mo, formerly Max 20x) includes $200/mo in API credits, and Team includes up to $500/mo pooled (Pro includes $0 API credits). That is an announced credit, not a seat-price change. Confirm in the Help Center article before you book it as included usage.

Frequently Asked Questions

How much does Claude API cost for a small project?

A small project making 100–500 API calls per day with Haiku 5.5 on the under-100k band is cheaper than the same volume on Haiku 4.5 ($1 / $5). Sonnet 5.5 at the same volume is cheaper than Sonnet 4.6 because the list is $2 / $10, not $3 / $15.

Is there a free tier for the Claude API?

Anthropic does not offer a permanent free API tier. You need to add a payment method and load credits. The October 7 Max and Team credit is a subscriber allowance, not a public free tier.

What's the cheapest way to use the Claude API?

Use Haiku 5.5 on prompts under 100,000 tokens ($0.10/MTok input), enable prompt caching, and batch non-real-time work (50% off). Haiku 4.5 remains $1 / $5 if a workload is pinned to that SKU. For mid-tier quality, prefer Sonnet 5.5 at $2 / $10 over Sonnet 4.6 at $3 / $15 unless you need the 4.6 SKU specifically. For Opus-class work, prefer Opus 5.5 at $4 / $20 over Opus 5 at $5 / $25 unless you are pinned to the older SKU.

How do Claude API costs compare to OpenAI?

This page's prior OpenAI pair was gpt-6.1-sol at $2 / $10 short-context and GPT-5 at $1.25 / $10. Those OpenAI rows were not re-verified on this pass. Sonnet 5.5 at $2 / $10 is unchanged. Haiku 5.5 at $0.10 / $0.50 under 100k is the new cheap Claude band. Compare the exact SKU pair, not the family name.

What happens when I hit a Claude API rate limit?

The API returns HTTP 429. Back off and retry with exponential backoff; if you hit limits regularly, request a higher spend tier in the Claude Console. Our Claude rate limits guide covers the RPM/TPM tiers and the full 429 playbook.

Do cached tokens cost the same as fresh input?

No. Cache reads run 0.025×–0.10× the input price depending on model — $0.10/MTok on Sonnet 5.5, $0.20 on Sonnet 5 and Opus 5.5, $0.20 on Fable 5.2 ($0.25 on legacy Fable 5.1), $0.01 on Haiku 5.5 under 100k, $0.10 on Haiku 4.5. On cache-heavy workloads the read line sets your bill, not the input line.

Related: Claude AI pricing guide · Fable 5 pricing, cache rates & plan access · Claude rate limits & tiers · How to get an Anthropic API key · How much does Claude AI cost

Primary & Reference Links Cited on This Page

- platform.claude.com pricing — `https://platform.claude.com/docs/en/about-claude/pricing`
- Haiku 5.5 — `https://www.anthropic.com/claude-haiku-5-5`
- Help Center article — `https://support.claude.com/en/articles/17154008`
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-api-token-meter-visual.jpg
- Tygart Media: Claude rate limits guide
- Tygart Media: Claude AI pricing guide
- Tygart Media: Fable 5 pricing, cache rates & plan access
- Tygart Media: How to get an Anthropic API key
- Tygart Media: How much does Claude AI cost