Claude LLM Quick Reference

Currently available models · June 2026

FLAGSHIP · NEWEST

Claude Opus 4.8

Released May 28, 2026

BEST FOR

Best for agentic coding, long-context reasoning, and production reliability. 4× less likely to miss code flaws. Most honest Claude model yet.

FLAGSHIP

Claude Opus 4.7

Released April 16, 2026

BEST FOR

Best for complex architecture decisions, large codebase analysis, agentic tasks requiring self-verification, and advanced vision workflows.

FLAGSHIP

Claude Opus 4.6

Released February 4, 2026

BEST FOR

Best for long-horizon autonomous tasks, multi-agent team workflows, and deep reasoning over very large contexts up to 1M tokens.

RECOMMENDED

Claude Sonnet 4.6

Released February 17, 2026

BEST FOR

Best for most professional tasks — coding, writing, analysis, and office workflows. Near-Opus quality at a fraction of the cost.

BUDGET

Claude Haiku 4.5

Released October 15, 2025

BEST FOR

Best for high-volume, cost-sensitive tasks — quick edits, file reads, simple Q&A;, customer service automation, and rapid scripting.

RESTRICTED PREVIEW

Claude Mythos

Invitation-only · 2026

BEST FOR

Defensive cybersecurity research only. Part of Project Glasswing — not publicly available, no self-serve access.

Quick pick: Use Sonnet 4.6 for most tasks. Step up to Opus 4.8 for best-in-class coding & reliability. Use Opus 4.6/4.7 for depth and autonomy. Use Haiku 4.5 when speed and cost matter most.

Token pricing · June 2026

MODEL	INPUT / M TOKENS	OUTPUT / M TOKENS	NOTES
Claude Opus 4.8	$5.00	$25.00	New Dynamic Workflows; 4× fewer missed code flaws
Claude Opus 4.7	$5.00	$25.00	New tokenizer — up to 35% more tokens vs 4.6
Claude Opus 4.6	$5.00	$25.00	1M context, no surcharge
Claude Sonnet 4.6	$3.00	$15.00	Best price-to-performance for most tasks
Claude Haiku 4.5	$1.00	$5.00	200K context window
Claude Mythos	—	—	Pricing not publicly disclosed

5x rule: Every model prices output at exactly 5× input. Know your input cost? Multiply by 5 for output.

DISCOUNT LEVERS — APPLY TO ALL MODELS

Prompt caching — up to 90% off input

Cache your system prompt, tool definitions, or static document context. First request pays a 25% write premium; every cache hit costs ~10% of standard rate.

Batch processing — 50% off all tokens

Submit async jobs via the Batch API. Results return within 24 hours. Best for pipelines, data processing, and workflows that don't need real-time responses.

Heads up: Opus 4.7 vs Opus 4.8 tokenizer costs

Opus 4.7 introduced a new tokenizer that produces up to 35% more tokens for the same input, raising effective costs even though the rate card is unchanged at $5/$25. Opus 4.8 reverses this — it uses ~35% fewer output tokens than 4.7, making it cheaper to run at the same list price. If you are on Opus 4.7, benchmark your workloads against 4.8: you may get better results for less.