ANTHROPIC

Claude Models

Which Model to Use and When

Claude Opus 4

Claude Sonnet 4

Claude Haiku 4.5

A Practical Decision Guide · 2025

The Claude Model Family

Three tiers intelligence, balance, and speed all built on Constitutional AI

Most Intelligent

Claude Opus 4

Frontier reasoning for the hardest tasks

- Complex multi-step reasoning
- Advanced coding & architecture
- Nuanced research & analysis
- Long-horizon agentic tasks
- Deep document understanding

Best Balance

Claude Sonnet 4

High capability at production speed

- Enterprise Q&A & summarization
- Code generation & review
- Customer-facing AI features
- RAG & document pipelines
- Creative & professional writing

Fastest & Cheapest

Claude Haiku 4.5

Real-time, high-volume, low-cost tasks

- Real-time chat & agents
- High-volume classification
- Low-latency API responses
- Data extraction at scale
- Mobile & edge use cases

Which Claude Should You Use?

Match your task to the right tier intelligence, balance, or speed

USE CASE / SCENARIO	MODEL	WHY IT FITS
Complex research & multi-step reasoning	Opus 4	Frontier-level intelligence for the hardest real-world problems
Advanced coding & system architecture	Opus 4	Deep reasoning across large codebases and architectural trade-offs
Enterprise Q&A & document summarization	Sonnet 4	Strong capability at production speed ideal cost/quality balance
Customer-facing AI products & chatbots	Sonnet 4	Reliable, nuanced responses with low enough latency for live use
Real-time chat & conversational agents	Haiku 4.5	Sub-second responses optimized for interactive, high-volume use
High-volume classification & extraction	Haiku 4.5	Lowest cost per token process millions of records affordably
Long-horizon agentic & tool-use workflows	Opus 4	Best multi-step planning, tool calling, and error-recovery logic

💡 Pro tip: Default to Sonnet 4 for most production workloads. Step up to Opus 4 for hard reasoning or agentic tasks. Use Haiku 4.5 to maximize throughput and minimize cost.