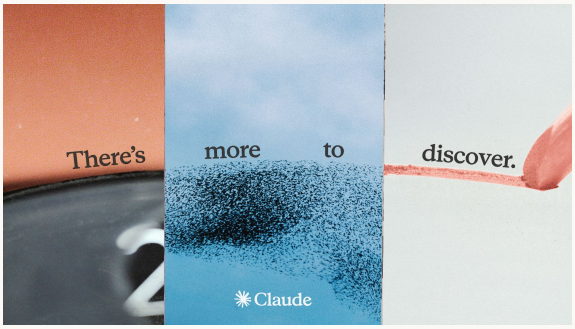
Claude BY ANTHROP\C

Haiku 5.5: Built for high-volume work



Claude Haiku 5.5 is the fastest and most efficient model in the Claude 5.5 family, built for high-volume, latency-sensitive agentic work.

It also costs around 75% less than Claude Haiku 4.5 for most tasks. For prompts up to 100K tokens, it costs $0.10 / $0.50 per million input / output tokens. Above 100K tokens, it costs $0.50 / $2.50.

Use it for:

Subagents

A fast, cost-efficient subagent for coding and well-defined tasks. Larger models like Opus 5.5 plan the work and hand off subtasks to Haiku 5.5.

Coding

Supports more complex coding agents by summarizing agentic transcripts, classifying and routing requests, and compacting context.

Knowledge work

Pulls key fields from small-to-medium documents and does first-pass reads to sort and flag what needs a closer look.

Interactive apps

Fast enough for live support and customer chat, so simple conversations get quick, helpful answers without a wait.

High-volume tasks

Handles classification, summarization, and text generation at a speed and cost that make processing large volumes of content practical.

Computer use

A computer use subagent for repetitive tasks like form filling, data entry, and moving information between apps.

When to use what model

Model	Use cases
Claude Fable 5.1	Frontier problems: long-running agents, net-new use cases, hardest coding & reasoning, full autonomy
Claude Opus 5.5	Your daily driver for complex and reasoning-intensive enterprise work as well as complex and production coding
Claude Sonnet 5.5	A cost-effective alternative to Opus for well-scoped work: building features, fixing bugs, short polished documents and decks, slides, and spreadsheets.
Claude Haiku 5.5	High-volume, latency-sensitive work: subagents, classification, summarization

Early users reported that Claude Haiku 5.5 is a clear step up from Claude Haiku 4.5 on customers' own evals, with faster responses. They tested it on the work where volume and latency matter most: document Q&A, routing, compaction, and the subagents inside their agent products. Several also noted how dependable it is. It catches its own mistakes and says plainly what it has and hasn't done instead of inventing an answer.

Why Update Your Agent To Claude Haiku 5.5

More capable

A step up from Haiku 4.5 across coding, tool use, computer use, and agents.

Cost efficient

The average Haiku 4.5 customer saves 75% on their Haiku bill.

Effort controls

The first Haiku with effort controls. Tune cost against intelligence for each task.

To update, change your model string to claude-haiku-5-5. The claude-api skill, pre-installed in Claude Code, can make the change for you with /claude-api migrate.

Claude Haiku 5.5

	Haiku 5.5	Haiku 4.5	GPT-6 Luna	Sonnet 5.5 For reference
Knowledge work GDPval-AA v2.1	**1620**	735	1437	1840
Knowledge work AA-Briefcase v1.1	**1578**	614	1336	1824
Computer use OSWorld 2.1 (Offline subset)	**72.4%**	15.7%	48.9%	83.9%
Multidisciplinary reasoning Humanity's Last Exam (no tools)	**45.9%**	10.2%	—	56.9%
Multidisciplinary reasoning Humanity's Last Exam (with tools)	**57.4%**	18.7%	—	64.5%
Agentic coding Terminal-Bench 4.0	**39.2%**	0.0%	16.4%	70.6%
Agentic coding FrontierCode 1.1 (Main)	**46.4%**	—	42.4%	52.1% Xhigh
Visual reasoning Chartography (no tools)	**46.4%**	6.4%	29.1%	61.6%

For details on how we run our evaluations, see the Haiku 5.5 System Card.