



CONSULTIPLY DAILY BRIEF • SMB
ANTHROPIC • OCT 7, 2026 • TOOLS

A cheaper, faster small model for the high-volume AI work your team already runs

ANTHROPIC
CLAUDE HAIKU 5.5

On Oct 7, 2026, Anthropic released Claude Haiku 5.5, calling it the cheapest, fastest, and most capable small model it has shipped. Anthropic positions it for summaries, classification, database queries, subagents, live support, and browser use, and says it costs about 75% less to run than Haiku 4.5 on average.

~75% less

Average cost to run vs Haiku 4.5 (Anthropic)

\$0.10 / \$0.50

Input per 1M tokens · prompts up to / over 100k

92.8%

HubSpot CRM eval suite score (vendor quote)

Available now

Claude Platform, AWS, Google Cloud, Azure

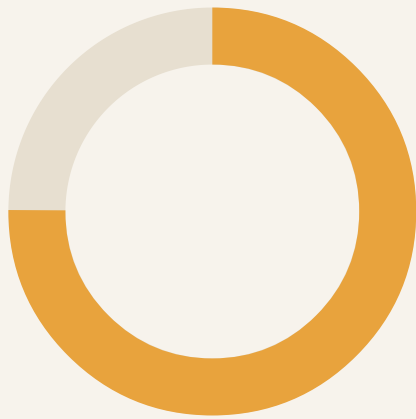
FIGURES FROM ANTHROPIC'S CLAUDE HAIKU 5.5 ANNOUNCEMENT, OCT 7, 2026 · VENDOR CLAIMS LABELED · NO PARTNERSHIP IMPLIED

OWNER TAKEAWAY

Keep Sonnet or Opus on hand, multi-step work. Route the repetitive volume (lookups, summaries, classification, subagents) to Haiku 5.5 and measure cost and latency on one workflow before you expand.

About 75% less to run than Haiku 4.5.

Anthropic prices Haiku 5.5 far below Haiku 4.5, especially on prompts up to 100k tokens (about 90% of prior Haiku requests). That is the number that matters when volume work piles up.



~75%

ANTHROPIC · OCT 7, 2026

less average cost to run vs Haiku 4.5

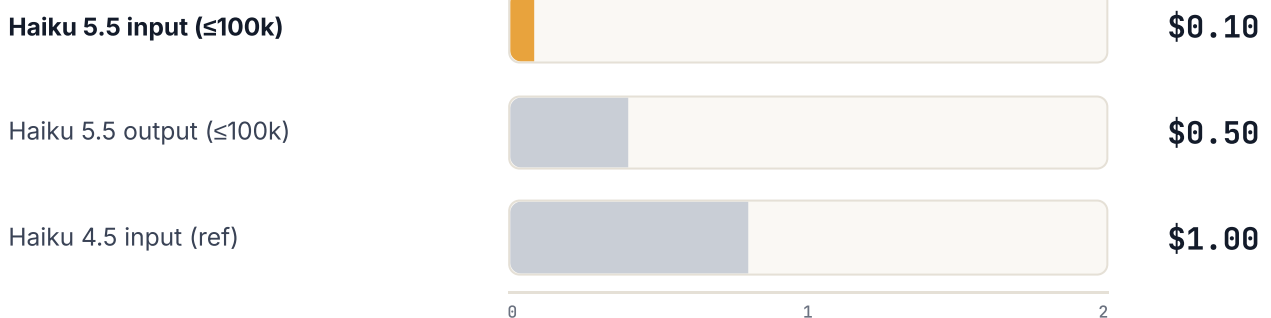
About 90% cheaper on prompts up to 100k tokens; 50% cheaper over 100k (Anthropic).

API ID · CLAUDE-HAIKU-5-5

AWS · GCP · AZURE

ADJUSTABLE EFFORT

Haiku 5.5 API price per 1M tokens (prompts up to 100k)



SOURCE · ANTHROPIC PRICING TABLE, CLAUDE HAIKU 5.5 ANNOUNCEMENT, OCT 7, 2026 · OVER-100K TIER IS HIGHER

FIT

High-volume narrow jobs

Summaries, classification, lookups, subagents.

SPEED

Fastest Haiku yet

Anthropic: fastest at standard speed; fit for live support and browser use.

PAIRING

Sidekick, not lead

Opus/Sonnet lead on hard coding; Haiku pulls the narrow pieces.

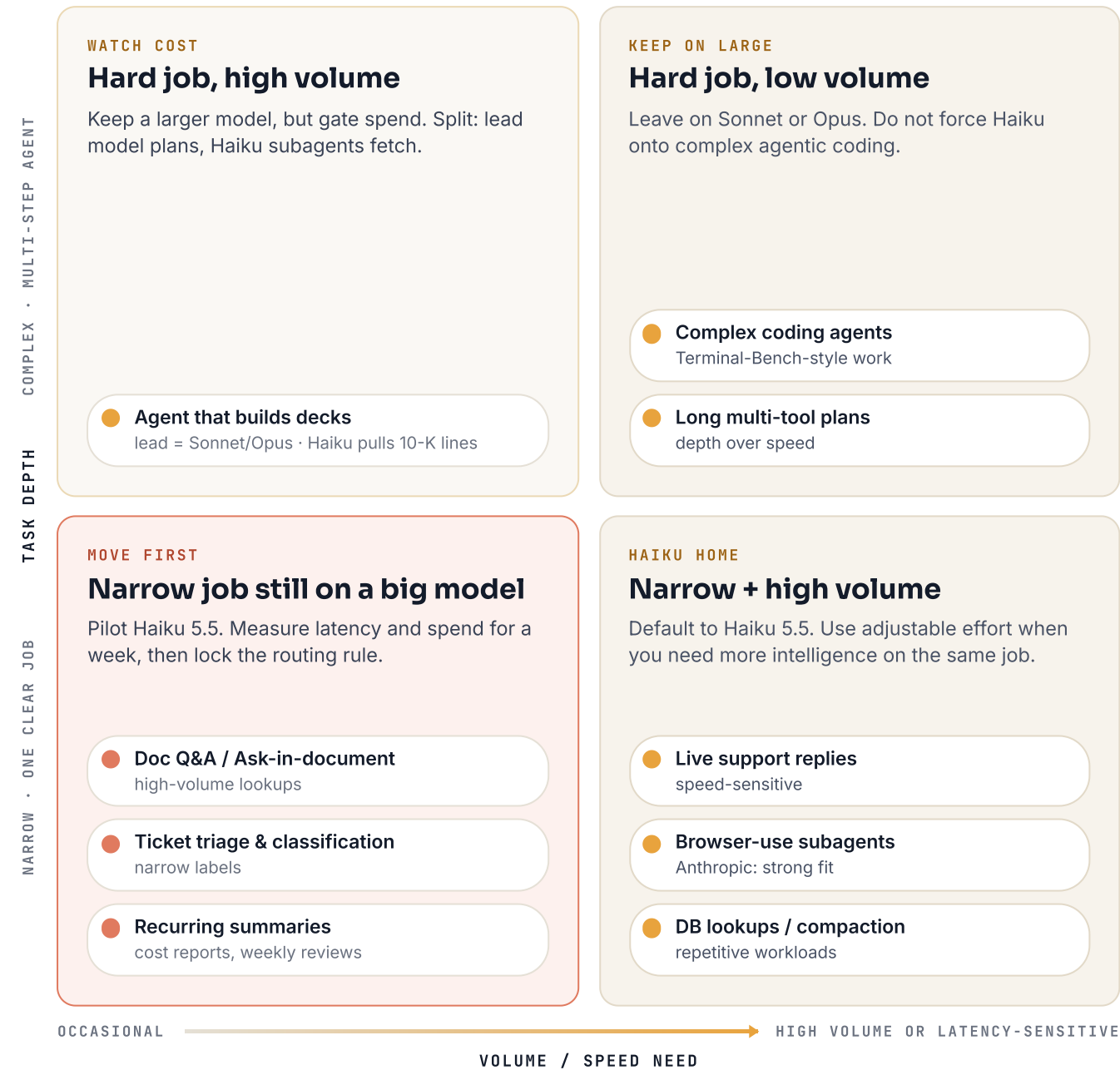
FIGURES AND QUOTES · ANTHROPIC CLAUDE HAIKU 5.5 ANNOUNCEMENT, OCT 7, 2026 · VENDOR CLAIMS LABELED

CUSTOMER SIGNAL (ANTHROPIC)

Asana: over 30% lower task latency and up to 2.5x faster inference per turn vs its current model. HubSpot: 92.8% on its CRM eval suite (Anthropic customer quotes).

Sort each AI job on volume and depth.

Anthropic says Haiku 5.5 fits narrowly scoped, high-volume work, while Sonnet 5.5 and Opus 5.5 stay better for complex agentic coding. Plot your workflows, then move the bottom-left box first.



ROUTING COLUMNS TO PUT ON ONE PAGE

JOB	VOLUME	DEPTH	MODEL NOW
What it does	Calls / week	Narrow · complex	Haiku · Sonnet · Opus

ROUTING MATRIX IS CONSULTIPLY GUIDANCE · FIT CLAIMS PER ANTHROPIC HAIKU 5.5 ANNOUNCEMENT, OCT 7, 2026

THIS WEEK

List three high-volume AI jobs in your business. Move one narrow job onto Haiku 5.5, keep complex agent work on a larger model, and name who checks cost and quality next Friday.



THIS WEEK • KEEP EXPLORING

Owner checklist: put Haiku 5.5 on the right jobs

- ☐ **List your high-volume AI tasks**
Summaries, classification, lookups, ticket triage, recurring report pulls. Note which model runs each one today.
- ☐ **Pilot Haiku 5.5 on one workflow**
Pick one narrow job. Compare latency and spend against your current model for a week of real traffic.
- ☐ **Keep a larger model on complex agent work**
Anthropic says Sonnet 5.5 and Opus 5.5 stay better for complex agentic coding. Do not force Haiku onto those jobs.
- ☐ **Name who watches cost and quality**
One person owns the routing rule and the weekly cost check. Expand only when both hold up.

SOURCES • CLICK TO OPEN

Anthropic • Introducing Claude Haiku 5.5 (Oct 7, 2026)

Primary: pricing, benchmarks, customer quotes, availability, effort setting, Sonnet cache-read cut.

anthropic.com/claude-haiku-5-5

Anthropic • Claude Haiku 5.5 System Card (Oct 7, 2026)

Evaluation details Anthropic links from the announcement for how benchmarks were run.

[anthropic.com](#) [CDN](#) • [Haiku 5.5 System Card PDF](#)

Want help routing models so volume work stays cheap without dropping quality?

consultiplify.ai