

Anthropic releases Claude Haiku 5.5 and halves Sonnet 5.5 cache-read prices

Source copy kept by wheresthe.ai for the story "Anthropic releases Claude Haiku 5.5 and halves Sonnet 5.5 cache-read prices" (<https://wheresthe.ai/d/anthropic-claude-haiku-5-5-release-2026-10/>).

Source: <https://www.anthropic.com/claude-haiku-5-5>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 55a70e85ced33f34eed9cf441bc5f353d8e637398d981898c3f4e3b571c97771

October 7, 2026 Claude Haiku 5.5 Skip intro Scroll down

Introducing Claude Haiku 5.5: the cheapest, fastest, and most capable small model we've ever released. Claude Haiku 5.5 is designed for high-volume, cost-sensitive tasks. It reliably handles quick and repetitive workloads (like summaries, compactions, database queries, and classification requests). It pairs well with Opus 5.5 and Sonnet 5.5 as a subagent on coding work. And, since it's also our fastest model to date, it works especially well for speed-sensitive tasks like live customer support and browser use.¹ Haiku 5.5 is available at a much lower price than Haiku 4.5. On average, it now costs around 75% less to run.² Along with this launch, we're making improvements to the value of our model range. We're halving the price of Claude Sonnet 5.5's cache reads, which means Sonnet 5.5 now runs around 20% cheaper on most agentic work. And we're introducing a new monthly API credit for our Claude Max and Team subscribers, designed to support our users in building new agents and applications that run on the Claude Platform. Performance Here's how Claude Haiku 5.5 performs across a range of benchmarks: Haiku 5.5 Haiku 4.5 GPT-6 Luna Sonnet 5.5 For reference Knowledge work GDPval-AA v2.1 Knowledge work GDPval-AA v2.1 1620 735 1437 1840 Knowledge work AA-Briefcase v1.1 Knowledge work AA-Briefcase v1.1 1578 614 1336 1824 Computer use OSWorld 2.1 Computer use OSWorld 2.1 72.4% Offline subset 15.7% Offline subset 48.9% Offline subset 83.9% Offline subset Multidisciplinary reasoning Humanity's Last Exam Multidisciplinary reasoning Humanity's Last Exam 45.9% no tools 10.2% no tools — 56.9% no tools 57.4% with tools 18.7% with tools — 64.5% with tools Agentic coding Terminal-Bench 4.0 Agentic coding Terminal-Bench 4.0 39.2% 0.0% 16.4% 70.6% Agentic coding FrontierCode 1.1 (Main) Agentic coding FrontierCode 1.1 (Main) 46.4% — 42.4% 52.1% Xhigh Visual reasoning Chartography Visual reasoning Chartography 46.4% no tools 6.4% no tools 29.1% no tools 61.6% no tools

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

Haiku 5.5 is our first Haiku-class model to come with an adjustable effort setting. This means that, as with our other models, users can decide whether to optimize for cost or intelligence. The charts below show how Haiku 5.5 performs on three benchmarks at each effort setting: Computer use: OSWorld Knowledge work: GDPval-AA Multidisciplinary reasoning: Humanity's Last Exam

Computer use: OSWorld Knowledge work: GDPval-AA Multidisciplinary reasoning: Humanity's Last Exam

OSWorld 2.1 (offline subset) Accuracy vs. cost Haiku 5.5 Haiku 4.5 Sonnet 5.5 GPT-6 Luna

OSWorld 2.1 measures how well agents can operate a real computer to finish long, multi-step tasks.

GDPval-AA v2.1 Accuracy vs. cost Haiku 5.5 Haiku 4.5 Sonnet 5.5 GPT-6 Luna

Artificial Analysis's GDPval-AA v2.1 evaluates agents on real-world professional work across 44 occupations.

Humanity's Last Exam (no tools) Accuracy vs. cost Haiku 5.5 Haiku 4.5 Sonnet 5.5

Humanity's Last Exam (HLE) is a test of expert-level academic knowledge and reasoning.

In early testing, our customers reported results consistent with the performance and cost improvements shown above. Here's what they told us about the new model: Asana HubSpot AlphaSense Box Rogo Cognition

Asana HubSpot AlphaSense Box Rogo Cognition

Quote "We're very impressed with Claude Haiku 5.5, particularly its speed. We ran it through our eval suite for AI Teammates, our AI agent product, covering use cases like triaging bugs, setting up projects, and searching large portfolios to surface high-risk or overdue work. Compared with the model we use today, we saw over a 30%

reduction in latency for task completions and up to 2.5x faster inference per agent turn. It's a noticeably snappier experience."

Company Asana Author Aaron Vinh, Staff Software Engineer

Quote "At HubSpot, we use simulated portals to evaluate new models on CRM tasks like reporting on deals. We mostly test the smaller, more efficient models, and Claude Haiku 5.5 got the best score we've seen on this suite yet, at 92.8% averaged over three runs. One CRM audit task asks models to identify stale but ambiguous records. Across all of the models we tested, Haiku 5.5 was fastest to complete the task, and had the highest hit rate and the lowest false positive rate."

Company HubSpot Author Ze'ev Klapow, Distinguished Software Engineer

Quote "Ask in Document is one of our big sources of spend, doing about 8M calls a week in production. It answers very specific questions on top of one or a few documents. We ran 400 queries, and Claude Haiku 5.5 was a statistically significant improvement over Haiku 4.5: 0.84 vs. 0.76."

Company AlphaSense Author Daniel Campos, Distinguished Engineer

Quote "Our customers use Box AI across large volumes of their enterprise content. With widespread usage comes the need to manage efficiency and cost, and to find the best model to suit the task at hand. In early testing, Claude Haiku 5.5 scored 11 points higher than Haiku 4.5 at about half the latency. We'd put it to use on analytical work that runs at scale, from cost reports to financial summaries and weekly recurring reviews."

Company Box Author Yashodha Bhavnani, VP of AI Products

Quote "The short and high-volume work is where Claude Haiku 5.5 fits for us, like quick lookups, subagents, and summaries. While a bigger model builds the deck, a Haiku 5.5 subagent goes into the 10-K and pulls the segment revenue line the deck needs. It's accurate enough that we'd trust it there, and fast and cheap enough that we can run it a lot."

Company Rogo Author Alex Wang, Applied AI

Quote "Claude Haiku 5.5 joins the sidekick lineup in Devin Fusion as an excellent option. With Haiku 5.5 as the sidekick, Fusion holds a top-tier FrontierCode score of 66.2 while cutting cost and latency. You can try it today in the Devin CLI with Opus 5.5 as the lead."

Company Cognition Author Walden Yan, Co-Founder & CPO

Pricing The table below shows how Claude Haiku 5.5's pricing compares to our other models. Haiku 5.5 is especially good value when used for tasks with prompts up to 100,000 tokens, which make up around 90% of requests to our previous Haiku model. Price per 1 million tokens Haiku 5.5 prompts up to / over 100k Haiku 4.5 Sonnet 5.5 Cache reads \$0.01 / \$0.05 \$0.10 \$0.10 Cache writes \$0.125 / \$0.625 \$1.25 \$2.50 Input tokens \$0.10 / \$0.50 \$1.00 \$2.00 Output tokens \$0.50 / \$2.50 \$5.00 \$10.00

Safety Alignment. Claude Haiku 5.5 shows major improvements across almost all of our alignment evaluations relative to Haiku 4.5. In particular, we found far fewer instances of misaligned behavior, and a lower willingness to cooperate with misuse. The model's system card describes our evaluation process and results in more detail. Safeguards. Consistent with its capabilities, Haiku 5.5's cybersecurity safeguards are more restrictive than Haiku 4.5's, but somewhat less restrictive than those we've applied to other recent models. In cybersecurity, they permit a wider range of defensive tasks than our safeguards for Sonnet 5.5, but they still block penetration testing and other techniques more likely to be used by attackers. Haiku 5.5's biology safeguards are the same as for Sonnet 5, Sonnet 5.5, and Opus 5. They allow research biology questions but restrict access to requests that we judge as likely to cause harm. Organizations working on wider-ranging biology and cyber activities can apply to our Life Sciences Verification Program and Cyber Verification Program . Availability Claude Haiku 5.5 is available now on all platforms, including Amazon Web Services, Google Cloud, and Microsoft Azure. On the Claude Platform, developers can get started with claude-haiku-5-5 . See our migration guide for details. Further updates Alongside our new pricing for Claude Haiku 5.5, we're making further improvements to the value of our models and products. First, starting today, we're lowering the price of cache reads on Claude Sonnet 5.5 . Cache reads now cost 50% less: \$0.10 per million tokens rather than \$0.20. Because cache reads make up a large share of models' token consumption, this reduces the cost of Sonnet 5.5 on most agentic tasks by around 20%. For instance, here's what the price cut means for Sonnet 5.5's performance relative to cost on Terminal-Bench 4.0: Terminal-Bench 4.0

Accuracy vs. cost Haiku 5.5 Haiku 4.5 Sonnet 5.5 (\$0.10 cache reads) Sonnet 5.5 (\$0.20 cache reads)

Terminal-Bench 4.0 measures how well a model can complete complex, multi-step professional tasks within a command-line interface.

This chart illustrates an important difference between Haiku 5.5 and our larger models. Sonnet 5.5 and Opus 5.5 remain better choices for complex agentic coding tasks like those measured by Terminal-Bench 4.0. By contrast, Haiku 5.5 is best suited to more narrowly scoped tasks that might otherwise have been cost-prohibitive with previous versions of Claude—like compaction, summarization, or subagent work. Second, this week, we'll roll out a new monthly API credit to all Max and Team subscribers for use on the Claude Platform . Max 5x users will get \$100 in credits per month, Max 20x users will get \$200, and Team subscribers will receive up to \$500, pooled across their users. These credits are designed to allow our users to experiment with building tools, apps, and agents that call our API. They can be used on any of our models. For more information, see our Help Center article . For developers, we're also updating our Claude Python and TypeScript SDKs to add support for computer use and browser use in beta. Haiku 5.5 is especially well-suited to these tasks, given its combination of speed, capability, and price. You can read more about this in our Claude Platform docs .

Footnotes 1 Claude Haiku 5.5 is our fastest model to date at each model's standard speed, although it runs less quickly than our Opus models in Fast Mode. 2 Claude Haiku 5.5 is priced 90% lower than Claude Haiku 4.5 for requests up to 100,000 tokens, and 50% lower for requests over 100,000 tokens. On Haiku 4.5, 90% of requests fell into the former category. This calculation also accounts for changes between Haiku 4.5 and Haiku 5.5 in how many tokens are used to complete a given piece of work: Haiku 5.5 has an updated tokenizer (similar to Sonnet 5.5's and Opus 5.5's), which means it uses slightly more tokens per task.