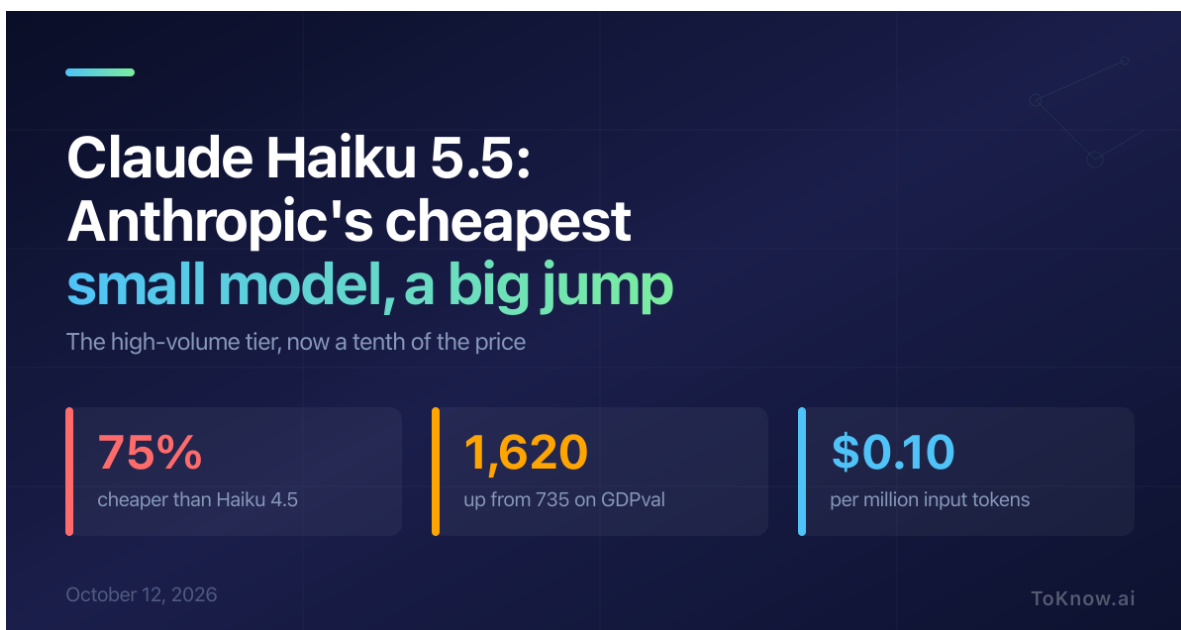


Claude Haiku 5.5: Anthropic's Cheapest Small Model Gets a Big Jump

Kabui, Charles

2026-10-12

[Read at ToKnow.ai](#)



Anthropic released Claude Haiku 5.5 on October 7, 2026, its cheapest and fastest small model, built for high-volume, repetitive work such as summaries, classification, routing, and database lookups, and to run as a helper beside the larger Sonnet 5.5 and Opus 5.5. It costs about 75% less to run than Haiku 4.5, the model it replaces. On GDPval-AA v2.1, a knowledge-work test, it scores 1,620 where Haiku 4.5 scored 735. On computer-use tasks it reaches 72.4% on

OSWorld 2.1 against 15.7%, and on multi-step coding 39.2% on Terminal-Bench 4.0 against 0.0%.

For prompts up to 100,000 tokens, Haiku 5.5 charges \$0.10 per million input tokens and \$0.50 per million output, roughly a tenth of Haiku 4.5. Independent tracker Artificial Analysis puts it at 43 on its Intelligence Index and 236 output tokens per second, among the fastest models it tracks. Anthropic also halved what Sonnet 5.5 pays to reuse old context, from \$0.20 to \$0.10 per million tokens, which trims about 20% off most long agent jobs.

Anthropic's new tokenizer counts the same text as about 30% more tokens than Haiku 4.5 did, so the average saving lands near 75% even though the per-token price is 90% lower on prompts under 100,000 tokens. Haiku 5.5 is also the first Haiku with an adjustable effort setting, so you can trade cost for thinking.

Read More: [Gemini 3.1 Flash Lite: Google's Fastest, Cheapest Frontier Model](#)

Sources:

- [Introducing Claude Haiku 5.5 \(Anthropic\)](#)
- [Claude Haiku 5.5 documentation \(Anthropic\)](#)
- [Claude Haiku 5.5: intelligence, performance and price analysis \(Artificial Analysis\)](#)
- [Anthropic releases Claude Haiku 5.5, cutting small-model API prices \(Unite.AI\)](#)
- [Claude Haiku 5.5 system card \(Anthropic\)](#)

Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. **Read more:** [/terms-of-service](#)