AI IN FINANCE | CLAUDE · AI COSTS · AUTOMATION · GLOBAL

Claude Haiku 5.5 cuts small-model prices by 90%: what it means for finance automation

Published 12 October 2026 · The Finance Intellect · Source-based summary

Anthropic released Claude Haiku 5.5 on 7 October 2026 at USD 0.10 per million input tokens, 90% below the previous Haiku model. High-volume finance tasks such as invoice coding and transaction classification become much cheaper to automate.

Global reference

- Vendor: Anthropic
- Topic: Claude Haiku 5.5, a small AI model for high-volume work
- Key date: Released 7 October 2026
- Availability: Claude Platform, Amazon Web Services, Google Cloud and Microsoft Azure

The pricing

- Prompts up to 100,000 tokens: USD 0.10 per million input tokens and USD 0.50 per million output tokens.
- Prompts over 100,000 tokens: USD 0.50 and USD 2.50.
- The previous model, Haiku 4.5, cost USD 1.00 and USD 5.00. Anthropic says the new price is 90% lower for requests up to 100,000 tokens and 50% lower above that.
- Anthropic says the 100,000-token threshold covers about 90% of requests made to the previous Haiku model.
- For comparison, Anthropic lists Sonnet 5.5 at USD 2.00 and USD 10.00.

What it is built for

- Summaries, classification, database queries, customer support, browser tasks and sub-agent work under larger models.
- It is the first Haiku model with an adjustable effort setting, which lets users trade cost against quality.
- Anthropic says its larger models, Sonnet 5.5 and Opus 5.5, remain better for complex agentic coding.

What it means for finance teams

- Illustration, on our own assumptions: coding 100,000 invoices at about 1,000 input tokens and 200 output tokens each would cost about USD 20 in model fees, against about USD 200 at Haiku 4.5 prices. This excludes integration, review time and any software licence.
- Model fees are no longer the main cost of high-volume automation. Review time, data quality and integration are.
- Good candidates: invoice coding, expense classification, bank transaction tagging, first-pass matching and document summaries.
- Poor candidates: accounting judgements such as provisions, impairment and revenue recognition.

What to do now

- Recalculate the cost per document in your AI business case.
- Route work by difficulty: a small model for routine items, a larger model or a person for exceptions.
- Test accuracy on your own data before you switch models, and keep sample checks after go-live.
- Keep prompts under 100,000 tokens where you can; the higher rate applies above that.

Our take

Each price cut moves the cost of AI from the model to the review. The design that wins is a cheap model for volume and a person on the exceptions. Spend the saving on better data and better checks.

Sources

- Anthropic: Claude Haiku 5.5 (7 October 2026)

Important. This summary is for general information and is not technology, legal or financial advice. It is based on public Anthropic information and may be incomplete or out of date. Check the original sources, and take professional advice, before acting.

ABOUT THE FINANCE INTELLECT

A Finance Knowledge & Resource Platform by Neocor Consultancy. Weekly tax and finance insights, Excel templates, calculators and a learning hub for finance leaders and professionals.

Platform: thefinanceintellect.com | **Owner:** neocorconsultancy.com