

Which Claude model for which task

Claude comes in three sizes, and most people quietly pay for the wrong one. Anthropic's new Sonnet 5 now handles work that needed their most expensive model six months ago, at a fraction of the price. Here is the whole picture on one page: what each model costs, what it is actually for, and the real tasks I run on each.

The 10-second version

Haiku for fast, simple, high-volume stuff. **Sonnet 5** for basically everything you do day to day. **Opus** only for the heaviest lifting. Match the model to the job and you stop overpaying for easy tasks, without losing quality where it counts.

What just changed

On June 30, Anthropic released Claude Sonnet 5. Their own words: it runs autonomously "at a level that, just a few months ago, required larger and more expensive models." Translation for the rest of us: the mid-size model now does most of what the flagship does. Opus 4.8 costs \$5 per million tokens in and \$25 out. Sonnet 5 costs \$3 and \$15, and until August 31 there is launch pricing at \$2 and \$10. That is 60% off flagship prices for near-flagship work.

How is that possible? Three things, no jargon. A smaller model burns less computing power per answer (you do not send a truck to deliver an envelope). The big model helps teach the small one, like a master training an apprentice. And the new models only think hard when the task actually needs it, instead of overthinking every little question.

The cheat sheet: how to pick, step by step

1. **Default to Sonnet 5.** Writing, coding, research, contracts, building agents: it is the everyday workhorse now. If you only remember one thing from this page, this is it.
2. **Drop to Haiku when the task is mechanical.** Sorting or tagging things, quick summaries, yes/no classification, anything you run hundreds of times. It costs \$1 in / \$5 out per million tokens, a fifth of Opus.
3. **Go up to Opus only for the heavy lifting.** Deep multi-step analysis, long overnight agent runs, the hardest problems where a wrong answer is expensive. That is what the \$5 / \$25 price is for.
4. **Switching is one click.** In the Claude app, use the model picker in the chat box. On the API, it is just the model name in your request. Nothing to install, nothing to migrate.
5. **Re-check every month or two.** This table changes fast. Six months ago the answer was different, and it will be different again. (That is half the reason this page exists.)

The three tiers at a glance

- **Haiku 4.5 · \$1 in / \$5 out per 1M tokens.** The small, fast one. Use for: sorting email, tagging leads, quick summaries, simple lookups, high-volume automations.
- **Sonnet 5 · \$3 in / \$15 out (\$2 / \$10 intro until Aug 31).** The everyday workhorse. Use for: writing and editing, coding, research, reading contracts, running agents, building documents.
- **Opus 4.8 · \$5 in / \$25 out.** The big brain. Use for: deep analysis, long-horizon agent runs, complex multi-step projects, the decisions you cannot afford to get wrong.

Real examples from a founder week

- **"Summarize these 40 support emails and tag the angry ones"** → Haiku. Mechanical, high volume, speed matters.
- **"Draft the investor update from these bullets"** → Sonnet 5. Quality writing, normal stakes.

- **"Read this 40-page contract and flag the risky clauses"** → Sonnet 5. It handles this fine now; this used to be an Opus job.
- **"Build and run the agent that reconciles a quarter of messy revenue data overnight"** → Opus. Long, hard, expensive to get wrong.

That is the whole trick: pick by job, not by habit. If this was useful, the newsletter is where I send the next one before it hits YouTube.

Built for curious people, not experts.

© 2026 Vasco Yaps. All rights reserved.