JULY 3, 2026

How to choose an AI model now that every lab ships three

The newest flagship is almost never the right default. The skill now is matching the model tier to the task, and most teams overpay by reaching for the top.

By **Dave Taylor**

Do you need the newest, most powerful model? Usually no. The first question in how to choose an AI model is not "which is smartest" — it's "what is this task actually asking of the model." A balanced tier like Claude Sonnet 5, at $3 per million input tokens against Opus 4.8's $5, lands near flagship quality on most real work at a fraction of the cost. Every major lab now ships three or four tiers on purpose. Reaching for the top of that ladder by reflex is the single most common way a team turns a working AI project into an expensive one. The model that fits the job beats the model that tops the benchmark, almost every time.

Do you actually need the flagship?

The flagship is built for the hardest 10% of your work, and you are probably not spending 90% of your budget there. Labs price the top tier for tasks that genuinely stretch a model — long multi-step reasoning, planning across many tools, the kind of problem where a weaker model quietly gets it wrong. Most business tasks are not that. Classifying an email, extracting fields from an invoice, drafting a first-pass reply, summarising a call — these are solved comfortably by a mid or fast tier, and the flagship's extra capability is spent on work that didn't need it. You pay flagship rates and get flagship-quality answers to questions a cheaper model would also have answered correctly.

The tell is your token bill against your task mix. If most of what your AI does is high-volume and repetitive, and you're running all of it on the top model, you're funding capability you never draw on. This is the same dynamic behind why AI token costs keep rising even as headline prices fall: teams default to the biggest model and let volume do the damage. The discipline is to start from the task and work up, not from the flagship and hope the budget holds. Ask what the job needs, then buy exactly that — no more, and rarely the most.

Every lab now ships a tier ladder

Every major lab ships a flagship-balanced-fast ladder now, and the price gap between the rungs is large. Anthropic's line runs from Claude Fable 5 and Opus 4.8 down through Sonnet 5 and Haiku 4.5: Opus 4.8 sits at $5 per million input tokens and $25 output, while Haiku 4.5 runs $1 and $5 — a 5x gap on both ends of the same family. Sonnet 5, the balanced tier, is $3 and $15, with an introductory $2 and $10 running through 31 August 2026, and Anthropic describes it as its most agentic Sonnet yet, near-Opus quality at a lower cost. That spread exists so you can put the right rung under the right task.

OpenAI's ladder has the same shape and a wider floor. Its current pricing runs from GPT-5.5 at $5 input and $30 output, through GPT-5.4 at $2.50 and $15, down to GPT-5.4 mini at $0.75 and $4.50 and GPT-5.4 nano at $0.20 and $1.25. The nano tier costs a twenty-fifth of the flagship on input. GPT-5.6 exists only in limited preview to select partners, so treat it as unavailable for planning. Google runs the same play with Gemini 3.1 Pro as the flagship and Gemini 3.5 Flash as the fast tier. Three labs, one pattern: a capable expensive top, a strong affordable middle, and a cheap fast bottom built for volume. The ladder is the product now, not the single best model.

Match the model to the task, not the headline

Match the model to the task, and the task tells you the tier — not the launch blog. Start by sorting your workload on two axes: how hard the reasoning is, and how much volume runs through it. High-volume, low-reasoning work — triage, classification, tagging, first-draft extraction — belongs on a fast tier like Haiku 4.5 or GPT-5.4 nano, where you're paying cents to process thousands of items and the quality is more than good enough. Low-volume, high-reasoning work — a legal analysis, a multi-step research task, a plan that has to hold together across a dozen tool calls — is where a flagship earns its price, because a cheaper model's mistakes cost more than the token savings.

The large middle is where most teams get it wrong, and where the balanced tier belongs. A model like Sonnet 5 or GPT-5.4 handles the bulk of real business work — drafting, summarising, answering grounded questions, moderate reasoning — at a quality gap from the flagship that a reader would struggle to spot, for half to a fifth of the cost. The headline announcing the newest top model tells you nothing about which rung your task sits on. A new flagship raises the ceiling; it rarely changes what the floor and the middle can already do. Choose by reading the job in front of you, not the release note that happened to land this week.

What "good enough" looks like in practice

"Good enough" is not a compromise — it's the correct answer when a cheaper model produces output you can't tell apart from the expensive one. Take invoice extraction: pull the vendor, date, total, and VAT from a PDF. A fast tier does this accurately across thousands of documents for a fraction of a cent each. Running that on a flagship changes nothing about the output and multiplies the bill. The right test is not "is this the best model" but "would a stronger model produce a materially better result for this specific task." If the answer is no, the cheaper model is not a trade-off. It's simply the right tool, and paying more buys you nothing but a smaller margin.

Where good enough breaks down is worth naming, because the line is real. A fast tier will drop threads on long reasoning chains, lose track across many tool calls, and get subtle judgement calls wrong in ways that read fluent and confident — the same trap we flagged in how a confidently wrong answer travels downstream before anyone checks it. So the practical method is to default to the cheapest tier that clears the bar for a given task, then step up only where output quality visibly suffers. You end up with a mix, not a single choice: most of the workload on cheap and fast, a slice on balanced, and the flagship held in reserve for the genuinely hard 10%. That mix is what a well-run AI system actually looks like.

How we pick models for a client agent

At Change by Design we build AI agents for SMBs, and we route by task rather than picking one model for the whole system. A single agent usually runs several models at once. The high-volume classification and triage step at the front — the part that decides what each incoming item is and where it goes — runs on a cheap fast tier, because it fires thousands of times and the decision is simple. A step deeper in the flow that has to reason across several tools, hold context, and make a judgement call runs on a stronger model, because that's where a cheap model's mistake would cost more than the savings. The flagship is rarely the default; it's the exception we reach for on the hard step, not the setting we leave running everywhere.

We build this on the same stack we work in daily — Anthropic and OpenAI for the models, n8n and Make for the orchestration that routes each step to the right tier. The payoff is direct: an agent that costs a fraction of what it would if every step ran on the top model, with no drop in the quality that reaches the client. Routing by task is not an optimisation you bolt on later. It's the design decision that decides whether an AI system pays back or quietly bleeds margin, and it's the first thing we settle before writing a line of the workflow.

The bottom line

The right way to choose an AI model in 2026 is to stop choosing one. Every major lab now ships a ladder — Anthropic's Opus 4.8 down to Haiku 4.5 at a 5x price gap, OpenAI's GPT-5.5 down to a nano tier at a twenty-fifth of the cost, Google's Pro and Flash — precisely so you can match the tier to the task instead of paying flagship rates for work a fast model handles cleanly. Sort your workload by reasoning difficulty and volume, default to the cheapest tier that clears the bar, and hold the flagship for the genuinely hard slice. The team that routes by task runs the same quality for a fraction of the spend. If you're building AI into your business and want to make sure it's routed to pay back rather than overpay, book a 30-minute working session and we'll map your workload to the right models.