Claude Model Cheat Sheet

Which Claude model to use, and why — the structure behind the lineup, so the logic still holds after the specific names and numbers change.

Prepared by Tienta, Inc. · Last reviewed August 2026
https://tienta.net/anthropic-model-sheet

START HERE

How to read this

Anthropic ships new models every few months and retires old ones on a rolling basis. Rather than pin exact specs here — they would be stale within weeks — this guide explains **how the model lineup is structured**, so the logic holds even after the specific names and numbers change, and links to the two pages Anthropic itself keeps current. Check those two first if a number below looks off.

Reference: Models overview (the master comparison table — pricing, context window, knowledge cutoff) · Choosing the right model (Anthropic's own decision framework)

PART 1

The naming pattern

This is the part that doesn't change often. Claude models come in **size tiers**, named after forms of writing, from fastest and cheapest to most capable.

Tier	Named for	General role
Haiku	Short form	Fastest, cheapest, highest-volume tasks
Sonnet	Mid-length form	The default workhorse — best balance of speed, cost, and intelligence
Opus	Long-form, most capable	Complex reasoning, coding, high-stakes work
Fable	Newer top tier	Anthropic's most capable widely released model, above Opus

Each tier gets versioned over time — a tier might go through several numbered releases in a year. **The tier name tells you the role; the version number tells you how recent it is.** When a client asks “should we use the new one,” the real question is usually “which tier does our task need,” not “which is newest.” A newer Haiku is not a substitute for Opus-level reasoning; it's just a faster, cheaper Haiku.

There is also a limited-access tier, **Mythos**, available only to vetted organizations for specific sensitive use cases such as defensive cybersecurity, through Anthropic's invitation-only Project Glasswing program. Not something most clients will encounter or need.

Reference: Model IDs and versioning

PART 2

The current lineup at a glance

Pull the live numbers from the Models overview table before sending anything to a client — pricing and context windows shift with each release.

Tier	Best for	Context window	Relative cost
Haiku	Real-time, high-volume, cost-sensitive tasks; simple classification, tagging, routing, first-pass drafts	Smaller	Lowest
Sonnet	The default choice for most production work — coding, data analysis, content creation, agentic tool use	Largest tier available	Mid
Opus	Complex agentic coding, enterprise work, multi-hour autonomous agents, large-scale refactoring, vision-heavy workflows	Largest tier available	High
Fable	The highest available capability — long-running agents, deepest reasoning, long-horizon agentic tasks, advanced research	Largest tier available	Highest

Anthropic's own rule of thumb, verbatim from their docs: “If you're unsure which model to use, start with Claude Opus [current flagship] for complex agentic coding and enterprise work. For workloads that need the highest available capability, use Claude Fable [current version].”

PART 3

Two ways to decide

Anthropic's framework offers two entry points, and they suit different situations.

3.1 Start efficiency-first

Begin with the cheapest and fastest tier, test the real use case, and upgrade only where there's a demonstrated capability gap. Best for prototyping, tight-latency applications, and cost-sensitive or high-volume straightforward tasks.

3.2 Start capability-first

Begin with the top tier, get it working well, then step down to a cheaper tier once the workflow is proven and optimized. Best for complex reasoning, scientific and mathematical work, advanced coding, and high-autonomy agentic work — anywhere accuracy matters more than early-stage cost.

3.3 The lever most clients don't know about

COMES UP ON CALLS

Several current models support an **effort parameter** that trades intelligence for speed and cost within the same model — often a better first move than switching tiers entirely. If a workflow feels slow or expensive on a capable model, try lowering effort before downgrading the model.

Reference: Effort · Choosing the right model

3.4 A third option: combine models

Route bulk and simple work to a cheap tier and escalate only the hard decisions to a frontier model, so most tokens are billed at the lower rate. Two common patterns:

- A cheap **executor** that asks a frontier **advisor** when it gets stuck.
- A frontier **orchestrator** that delegates bulk work to cheap **workers**.

Reference: Optimizing for cost and intelligence

PART 4

Where model choice actually happens

This connects back to the Admin Cheat Sheet — model selection shows up differently depending on which product a client is using.

Product	Where model choice lives	Who controls it
claude.ai chat	Model picker in the chat UI	The individual user, per conversation. Every paid plan can access all current tiers — the plan controls usage volume, not which models are available.
Claude Code (Team/ Enterprise)	`/model` in the CLI, or an org-wide default set by an admin	Org admins can set a default model for Claude Code users at `claude.ai/admin-settings/claude-code`, org-wide or per custom role.
Claude API (platform.claude.com)	The `model` field in each API request, referencing a specific model ID such as `claude-opus-5`	The developer, per API call. This is the one place model IDs matter to the byte, since a pinned ID is what gets billed and what determines behavior.

One nuance worth flagging to clients: on the API, model IDs are **pinned snapshots**, not evergreen pointers. `claude-opus-5` won't silently start meaning a different model next year the way "the current flagship" might in conversation. That's a feature for production stability, but it also means client integrations need a deliberate migration step when a new tier version ships — Anthropic publishes a migration guide for exactly this.

Reference: Model configuration (Claude Code) · Model deprecations · Upgrade between model versions

PART 5

Common client questions

Client says…	Point them to…
"Which model should we default to?"	The Sonnet tier, for nearly everyone — it's Anthropic's own "best combination of speed and intelligence" framing. Reserve Opus and Fable for specific hard tasks, not as the default.
"Our API bill is high — can we use a cheaper model?"	Try lowering the effort parameter first, then consider a mixed executor/advisor setup before a blanket downgrade — see §3.4.

Client says…	Point them to…
“Do we need to keep upgrading to the newest model?”	Not automatically — pinned model IDs mean nothing breaks by staying put. But check Model deprecations for retirement dates.
“What's the difference between claude.ai and the API?”	Same models; a different mechanism for choosing them (picker versus explicit ID) and different billing (per seat versus per token). See the Admin Cheat Sheet.
“Can non-technical staff pick the ‘best’ model themselves?”	Yes, in the claude.ai chat picker — every paid plan includes all current tiers. Consider an organization instruction reinforcing “Sonnet is the default, Opus is for X” so people aren't guessing per conversation.

PART 6

Canonical link directory

Bookmark these rather than quoting numbers from them — the pages stay current on their own.

Choosing and comparing models

Reference: Models overview (master comparison table) · Choosing the right model · Optimizing for cost and intelligence · Pricing

Versions, IDs, and migration

Reference: Model IDs and versioning · Model deprecations · Upgrade between model versions

Configuration and reference

Reference: Effort parameter · Model configuration (Claude Code) · Model cards (safety and capability documentation per model)

Notes on keeping this current

- The **lineup table** — tier names, pricing, context windows — is the part that goes stale fastest, since Anthropic ships new versions every few months. Re-pull it from Models overview before any client-facing send, or link the page live instead of quoting numbers.
- The **naming pattern and decision framework** — the Haiku, Sonnet, Opus, and Fable roles, efficiency-first versus capability-first, effort as a lever — have held steady across multiple model generations and are the durable part of this guide.
- If a client asks about a model name you don't recognize, check Models overview before guessing. Anthropic sometimes introduces a new top tier — Fable was one such addition — rather than just incrementing version numbers.