RESOURCE

How To Think About Model Selection In Claude

Model selection is part of understanding AI tokenomics: different models and reasoning levels use different amounts of computational capacity to complete a task. Using the most powerful option for every request can waste resources, while using too little reasoning can produce weaker results and require more iteration. The goal is to use the **lowest level of capability and reasoning that reliably produces the result you need**.

Although this resource focuses on Claude, the same way of thinking applies across AI platforms. ChatGPT exposes reasoning effort settings, Gemini offers adjustable thinking levels, and Copilot distinguishes between faster responses and deeper reasoning modes. Learn to recognize when a task needs speed, everyday reasoning, or deeper analysis, and you can make better model choices regardless of the platform.

Choosing Haiku, Sonnet, Opus, and Fable

If you are unsure which model to use, start with Opus 5. Move up or down the scale when the task asks for it.

LEAST CAPABILITY NEEDED — MOST

Haiku 4.5	Sonnet 5	Opus 5	Fable 5.1
Fastest, near-frontier	Speed and intelligence	Start here for most work	Demanding, long-horizon work

WHERE EACH MODEL SHINES

Fable 5.1

DEMANDING REASONING

Long-horizon agentic work, or when evals on Opus 5 at higher effort still fall short.

- Competition-level coding challenges
- Mathematical calculations and proofs

Slower · 1M context · thinking always on

Opus 5

THE DEFAULT CHOICE

Complex agentic coding and enterprise work. Start here for most workloads.

- Multi-step technical problems
- Comprehensive project planning

Moderate · 1M context · thinking always on

Sonnet 5

SPEED WITH INTELLIGENCE

The best combination of speed and intelligence for work you run often.

- Drafting and editing with context
- Detailed document analysis

Fast · 1M context · thinking optional

Haiku 4.5

CLEAR, REPEATABLE TASKS

The fastest model, for high-volume tasks where you know what good looks like.

- Extraction and classification
- Structured summaries at volume

Fastest · 200K context · thinking optional

Pick a thinking effort

Each model has a recommended effort level, marked as “Default” in the menu.

Low and Medium	Work well for routine tasks and stretch your usage further.
High	Offers the best overall balance of quality and speed.
Extra high	Designed for long-running coding and agentic tasks, offering deeper reasoning than high without the full token cost of max. Available on Opus 4.7 and newer models.
Max 3.5×	The most thorough option, best for tasks requiring the deepest possible reasoning and most thorough analysis.

Effort and thinking are separate settings. Effort controls how thorough Claude is; the thinking toggle controls whether Claude works through its reasoning in an expandable section first. Thinking cannot be turned off on Fable 5.1 or Opus 5. Max effort draws on usage limits well above the standard rate.