How to choose the right Claude model: Haiku, Sonnet, or Opus

Claude has three versions - Haiku, Sonnet, and Opus. Each version is designed for different types of work and uses your token limit in a different way.

Claude has three versions – Haiku, Sonnet, and Opus. Each version is designed for different types of tasks and uses your token limit differently. Using Opus for a task that Haiku can handle will waste your tokens without any benefit and slow down your workflow. In this guide, you will learn how to choose the most suitable Claude version to complete your work and help you use your token limit more comfortably.

The versions you get and the token limits depend on your Claude account package. The free package includes Haiku and Sonnet; the Pro and Max packages add more Opus and more limits.

Learn about 3 versions of Claude

Claude has three versions. Think of them as different tools designed for different tasks.

Model	Speed ?? limit	Most suitable for
Haiku 4.5	lightest	Quick answers, concise summaries, and simple extraction – whatever you want to do immediately.
Sonnet 4.6	Medium	Programming, writing, analysis, and multi-step workflows - the default all-purpose tool.
Opus 4.6	Heaviest	In-depth research and complex reasoning truly require persistent thinking.

Haiku is fast and lightweight. Haiku 4.5 is built for everyday needs, and its reasoning capabilities are on par with the Sonnet 4.0 model. When you need quick answers to simple questions, basic summaries, or synthesis, Haiku will do the job instantly. It's also most effective within your speed limits.

Sonnet is an everyday tool. Sonnet 4.6 delivers powerful reasoning capabilities for the kind of work you do every day—programming, writing, analyzing, researching, and solving complex problems. It's responsive enough for real-time collaboration and robust enough to handle most issues. It also handles computer usage, visual tasks, document creation, and spreadsheets well—making it a versatile default tool across a wide range of jobs. If you're unsure which model to choose, start here.

Opus is an expert in large-scale reasoning. Opus 4.6 is particularly excellent for specialized, complex tasks requiring advanced reasoning capabilities. It's built for real problems that require deep thinking over extended periods. It uses more of its speed limit, so you should reserve it for truly essential tasks. Opus is available on Pro packages and above.

Understanding speed limits

The rate limit dictates the number of tokens you can use within a given time period. Token consumption rates vary: Haiku is the lightest, Sonnet is medium, and Opus uses the most because it involves deeper reasoning.

Simpler models like Haiku are more "efficient" within your limits – you can ask more questions with fewer tokens. Opus uses more tokens because it performs more arguments. But if you use Opus for a task that Haiku can handle, you're unnecessarily using up your limits.

Both Sonnet 4.6 and Opus 4.6 feature more adaptive extended thinking, meaning Claude automatically adjusts its reasoning depth to suit the problem. Simple questions receive quick answers; more complex questions require more thought. This makes both models more token-efficient than previous versions – even with extended thinking enabled, easy questions don't consume as many tokens as before.

When should each model be used?

To strike a balance between speed limits and efficiency, it's best to consider the type of output you desire from Claude and choose the appropriate model.

Use Haiku for:

1. Simple, easy-to-understand questions with concise answers.
2. Tasks that only require quick lookup or categorization.
3. Extract specific information from text.
4. A simple summary or overview
5. Anything you want to do immediately without complicated reasoning.

Use Sonnet for:

1. Writing and creating content
2. Programming tasks - debugging, writing new code, refactoring
3. The analysis requires reasoning but shouldn't be overly complex.
4. Chatbots provide customer support with context and nuance.
5. Multi-step issues and workflows
6. Most "multipurpose" jobs leave you unsure which model to use.

Use Opus for:

1. In-depth research and analysis
2. Complex reasoning requires multiple steps.

3. Specialized tasks where accuracy is critical.
4. The problems you've been testing with Sonnet and the difficulties it's encountering.
5. Long documents require a high level of concentration to understand.

What do these models look like in reality?

Let's see how these patterns play out with a few common tasks.

Task	Select	Reason
Debugging code	Sonnet	Superior programming capabilities, rapid response, and clear error identification.
Summary of articles	Haiku	Simple content extraction, no complex reasoning required.
Analyzing complex research papers	Opus	In-depth analysis of lengthy specialized literature, including methodological critique and forward-looking insights.

It's worth noting that a new version of Claude isn't always as good as the old one. Each release is a separate training process, meaning a task that works well with Opus 4.5 might be better suited for Sonnet 4.6, or vice versa.

This balance can change with each release, so you should double-check your assumptions when a new version comes out. Take a few minutes to run your frequently used tasks on several different models. If you're using the free plan, upgrading will give you access to all three models and a higher speed limit. When a new model is released, you'll have more room to experiment.