

EVERYONE · 5 MINUTE READ

# Choosing a model and a thinking mode

Pick the smallest, fastest model that does your task well, switch on a thinking mode only for multi-step work, and test on your own tasks.

## What it is

Behind every AI assistant, such as Claude, ChatGPT, Microsoft Copilot or Gemini, sits a **model**: the trained system that reads your request and writes the answer. Many assistants let you choose between several models. Larger ones are more capable; smaller ones are faster and cheaper to run.

Many assistants also offer a **thinking mode** (sometimes called extended thinking or reasoning). The model works through the problem step by step before answering. It is slower, but often better on work with many steps.

Many assistants also offer an automatic setting that decides for you; you can usually override it, and some let you choose how much thinking to use.

Think of choosing who to ask at work: you would not book a senior analyst to fix a typo, or give a three-year budget plan to a quick helper.

Model names change often, so this guide describes tiers rather than products: a **fast tier**, a **balanced tier** and a **most capable tier**, each with or without thinking switched on.

## Why it matters

- The right choice gives a better answer, faster.
- Larger models and thinking modes use more of your **usage limit**, and in pay-per-use setups they can cost more per use.
- A thinking mode can make simple jobs slower without making them better.
- A weak model on a hard task can give confident but wrong answers, so the human check matters even more.

## How to do it

1. **Decide** whether the task is simple drafting or needs several steps of reasoning (analysis, maths, planning, comparing options).
2. **Start** with the fast or balanced tier for everyday writing, summarising and rewording.
3. **Switch on** a thinking mode for multi-step work: checking calculations, planning, weighing options against rules.
4. **Move up** to the most capable tier only when the smaller one gets things wrong or misses important points.
5. **Test** on your own task: give two models the same prompt and the same made-up data, then compare.
6. **Note** which choice worked for which job, so you and your team do not have to repeat the test.
7. **Check** the answer yourself, whichever model you used.

## Example

Two jobs, two choices.

**Job 1: a polite reply to a meeting request.** A fast model writes a good draft in seconds. Thinking mode adds nothing useful.

Draft a short, friendly reply accepting the meeting on [day] at [time]. Ask for the agenda in advance. Under 60 words.

**Job 2: comparing three supplier quotes against your purchasing rules.** This needs several steps: read each quote, apply each rule, total the costs, flag exceptions. Use the balanced or most capable tier with thinking switched on.

Here are three supplier quotes (made-up for this test) and our purchasing rules. For each quote:

1. Check it against every rule and say which it meets or breaks.
2. Show the total cost and how you calculated it.
3. Recommend one, with two reasons.

If any figure is unclear, say so rather than guessing.

You then recheck the totals and rule checks before anyone approves a purchase.

## Common mistakes

- **Always using the biggest model.** Fix: start small and move up only when the answer is not good enough.
- **Leaving thinking mode on for everything.** Fix: use it for multi-step work; switch it off for quick drafting.
- **Testing on someone else's example.** Fix: test on a real task of yours, using made-up or non-sensitive data and only approved tools.
- **Writing model names into team guidance.** Fix: describe the tier and the job ("fast tier for email replies"), because names change.

## Try it this week

1. Pick one simple task (an email reply) and one harder task (a short analysis with a few numbers), using made-up data.
2. Run each on the fast tier without thinking and note the result (5 minutes).
3. Run each again with a larger tier or thinking switched on (5 minutes).
4. Compare: was the slower answer actually better? Write one line per task saying which setting you will use next time (5 minutes).

## Going further

- **Tokens** are the small pieces of text a model reads and writes, roughly a short word or part of a longer word. Limits and prices are usually counted in tokens, so long documents and long answers use more.
- Thinking modes use extra tokens for the working (often hidden or shown only as a summary), which is why they count more against your limit.
- For repeated tasks at scale, ask IT about the cost per use of each tier before choosing a default.
- Re-test when a new model arrives (see guide 27).

## Key words

- **Model:** the trained AI system that reads your request and writes the answer.
- **Thinking mode:** a setting where the model works step by step before answering; slower, often better for complex tasks.
- **Token:** a small piece of text used to measure how much a model reads and writes.
- **Usage limit:** how much you can use an assistant in a given period under your plan.
- **Tier:** a group of models of similar speed, cost and capability.