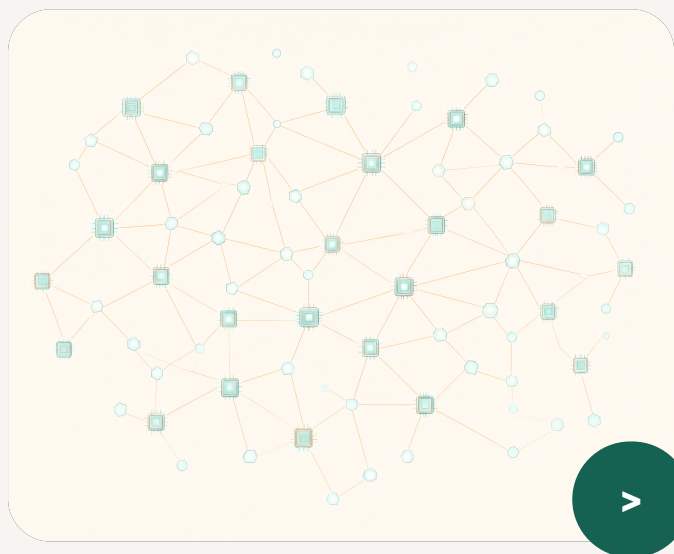


Which AI should you actually use?

There are now dozens of AI models, all claiming to be the best. This short, plain-English guide shows you how to choose the right one for the job — and how to test them yourself, in an afternoon, without any technical knowledge.



What you will be able to do

Tell a frontier model from an open one and a local one. Read a benchmark without being fooled. And pick the right model for each task — documents, numbers, writing, coding or confidential files.

Why the model you choose matters

Two AI tools can look identical and behave completely differently. One reads a 200-page report without losing the thread; another forgets the beginning by the end. One is private enough for confidential files; another is not. Choosing well saves time, money and, occasionally, your reputation.

WHAT IS “MODEL”?

The “engine” behind an AI tool. ChatGPT, Claude and Gemini are brands; each is powered by a model, and the model changes every few months.

WHAT IS “FRONTIER MODEL”?

The most capable (and most expensive) models, made by a handful of large companies. Best for hard reasoning and long documents.

WHAT IS “OPEN MODEL”?

A model whose weights are published, so anyone can download and run it — including on their own computer.

WHAT IS “LOCAL MODEL”?

An open model running on your own machine, not on a company's website.

“Best” is not a property of a model. It is a match between a model and a job.

■ IN SUMMARY

The right question is never “which is best?” but “which is best for this task, at this cost, with this data?”

The 2026 landscape, in plain English

Ignore the hype and there are really four families of model. Knowing which family you are dealing with tells you most of what you need.

The frontier leaders

The Claude Opus 5 / Sonnet 5 models, the GPT-6 models and the Gemini 3.8 Flash line. Strongest at hard reasoning and long documents; the most expensive; used through a website or an API.

The open-weight wave

Qwen 3.8, DeepSeek V4.1, GLM 5.3, Kimi K3 and Meta Muse Spark. Almost as capable as the frontier, free to download, and runnable on your own hardware.

Local models

Open models running on your own laptop or desktop — private by design, but limited by your machine's memory.

Chinese labs

Qwen, DeepSeek, Kimi and GLM are among the strongest open models in the world. They are widely used — but read the data question on the next pages before using them on sensitive work.

In 2026 a 27-billion-parameter open model — runnable on a single good graphics card — now performs at the level of the frontier models of a year earlier. The gap between free and paid is far narrower than people assume.

How to read a benchmark — without being fooled

A “benchmark” is a standard test that models are scored on, like an exam. They are useful — and easy to misread.

Benchmark	What it measures, and its catch
MMLU-Pro	General knowledge across subjects. Easy to over-train on; a high score does not mean good at your job.
GPQA Diamond	Hard science questions. Favours models with long “thinking” phases.
SWE-bench Verified	Fixing real software problems. The best proxy for agentic coding, but not for writing.
Arena / preference	How pleasing answers seem to voters. Rewards confident style — which is exactly what we must distrust.

■ KEY TAKEAWAYS

- Three ways scores mislead: the test may be in the training data; providers choose the flattering test; and a test is not your job.
- A model that wins a leaderboard can still invent a citation in your report.
- Always confirm with your own small test — shown next.

Build your own benchmark (the only one that counts)

You can rank models on your own work in about an hour, using the same idea as the handbook's "golden set".

1

Pick ten real tasks

Ten things you actually do — summarise a report, build a table from a survey, draft a section, translate a paragraph.

Why this matters: Your job is the only benchmark that matters.

2

Note the correct answer

One line each describing what a good answer must contain, and the mistake to watch for.

Why this matters: This is your mark scheme.

3

Run three or four models

Try a frontier model, a cheap one, an open one, and — if you have the hardware — a local one.

Why this matters: You will be surprised how close they are.

4

Score and compare

Mark each answer out of ten against your scheme, and note the time and cost.

Why this matters: Now you can choose by evidence, not reputation.

■ IN SUMMARY

One hour gives you a personal, un-riggable ranking of the models for your work.

■ GO FURTHER — THE PRO EDITION

The Pro edition gives the full procedure, a scoring sheet, a cost calculator and a hardware guide for running models locally.

\$79 · downloadable PDF with every template. Or get it included with the video course. For teams, write to contact@evaluationai.com.

One key, every model

The easiest way to try many models without many accounts is a “router” service such as OpenRouter. You sign up once, get one key, and can call almost every model — cheaply, paying only for what you use.

1

Create a free account

Go to openrouter.ai and sign up with an email or a Google account.

Why this matters: One account replaces a dozen.

2

Make one API key

In the dashboard, click “Create key”. Copy it somewhere safe — treat it like a password.

Why this matters: This single key unlocks every model.

3

Add a little credit

Top up a small amount, for example ten dollars. Most tasks cost a fraction of a cent.

Why this matters: No subscription; you pay per use.

4

Point your tool at it

Paste the key into any app that supports OpenRouter — including many chat apps and our own harness.

Why this matters: Now you can switch models with one word.

WHAT IS “API KEY”?

A long secret code that identifies you to the service. Never share it or paste it into a public place.

Which model for which task

A starting point. Confirm it with your own ten-task test, because the answers move every few months.

Your task	Reach for...
Long document analysis	A frontier model with a very large context window.
Extracting data into a table	A cheap, fast model with strict “structured output”.
Drafting and rewriting	A mid-priced model — a frontier model is overkill.
Numbers and statistics	A frontier or strong reasoning model; never trust the working.
Coding and spreadsheets	A coding-specialist model (for example a Qwen-Coder family).
Translation	A multilingual model; test on your own language pair.
Confidential client files	A local model, or a paid account with training switched off.

Rule of thumb: use the cheapest model that passes your ten-task test. Spend on the frontier only when the cost of an error is higher than the cost of the model.

A simple privacy traffic-light

Before you paste anything into any model, decide which light applies.

Green — public tool is fine

Material that is already public: published reports, public data, generic questions.

Amber — use your own paid account

Ordinary client work, anonymised. A paid account with “model training” switched off.

Red — local, or not at all

Named individuals, personal data, anything privileged or confidential. Use a model on your own machine, or don't paste it at all.

WHAT IS “TRAINING SWITCH”?

A setting in a paid AI account that stops your text being used to improve the company's models. Always turn it off for client work.

■ IN SUMMARY

The single safest habit: redact names and places first, whatever model you use.

Local models, honestly

A local model runs on your own computer. It is the most private option — and the most fiddly. Here is the plain truth about the trade-offs; the Pro edition covers the hardware in detail.

The good

Private by design. No subscription. Works offline. No per-word cost. A capable 8-billion model runs on a modest laptop.

The catch

Needs a machine with enough memory. Slower than the cloud. Setup and updates take patience — this is where most people stop.

If local is the right answer for you, do not struggle alone — we set it up for organisations. Write to contact@evaluationai.com.

■ KEY TAKEAWAYS

- A laptop with 8–16 GB can run small models; a good 24 GB graphics card runs a strong 27-billion model.
- A modern Mac with 32–64 GB of memory runs larger models quietly.
- The Pro edition includes a full hardware guide and a step-by-step local setup.

The words you'll meet

Token	A small chunk of text — roughly $\frac{3}{4}$ of a word. Models charge and measure in tokens.
Context window	How much text a model can hold in mind at once, measured in tokens.
Parameters	The size of a model, in billions. Bigger is not always better, but it usually needs more memory.
Quantisation	Shrinking a model to use less memory — at a small cost in quality.
MoE	“Mixture of experts”: a large model that only uses part of itself for each word, so it is fast.
Agent	A model that can take several steps and use tools on its own, rather than answering once.
RAG	Giving a model your own documents to read before it answers, so it relies on your sources.
Structured output	Asking a model to reply in a fixed shape — for example a table — so it can be used reliably.

■ WHERE TO GO NEXT

Pick well. Test for yourself. Stay private.

The free edition gives you the map. The Pro edition gives you the full hardware guide, the task-by-task router, the cost calculator and a step-by-step local setup.

Free edition — this guide

The four families, benchmarks, your own test, and the cheat sheet.

Pro edition — \$49

The complete model-selection system, with hardware tables and setup procedures.

evaluationai.com · contact@evaluationai.com