

THE LANGUAGE OF AI

THE BIG IDEA

One colleague says AI was useless; another says it saves hours every week. Very often they weren't using the same thing: a free account with a lighter model, versus a paid or campus account with the full lineup in a different app.

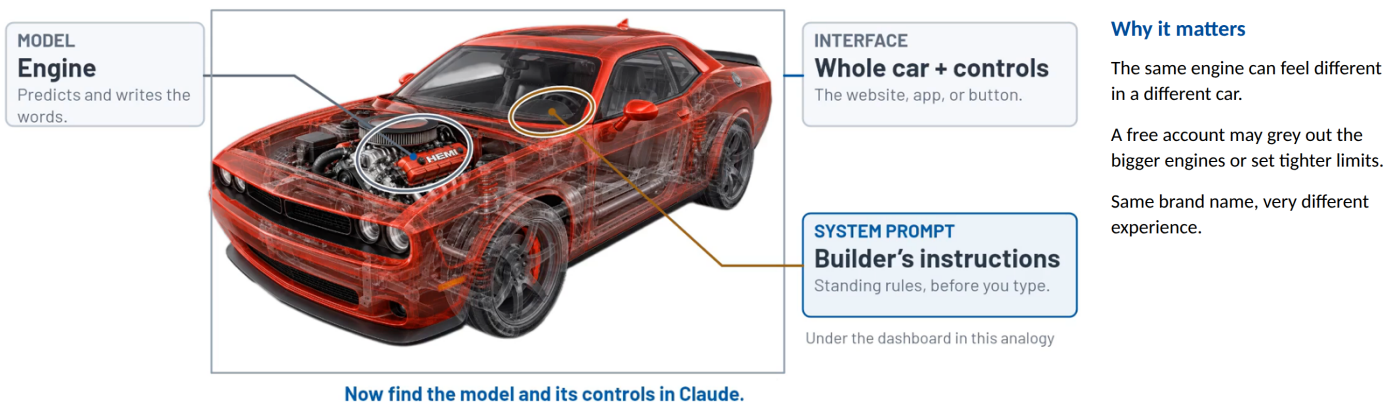
Same brand name, very different experience. Telling the model from the interface, knowing the three kinds of models, and asking four questions make the landscape a lot less confusing.

MODELS AND INTERFACES: THE ENGINE AND THE CAR Video 1:36

The **model** is the engine: the underlying AI system that predicts and writes the words. The **interface** is the car built around it: the website, the desktop app, the phone app, or a button inside a program you already use. Each company names its engines. When the video was made, Claude's lineup ran from Fable 5.1 (the most capable) and Opus 5.5, through Sonnet 5, to Haiku 4.5 (the most economical). Expect those names to change.

MODELS AND INTERFACES

One car. Three different jobs.



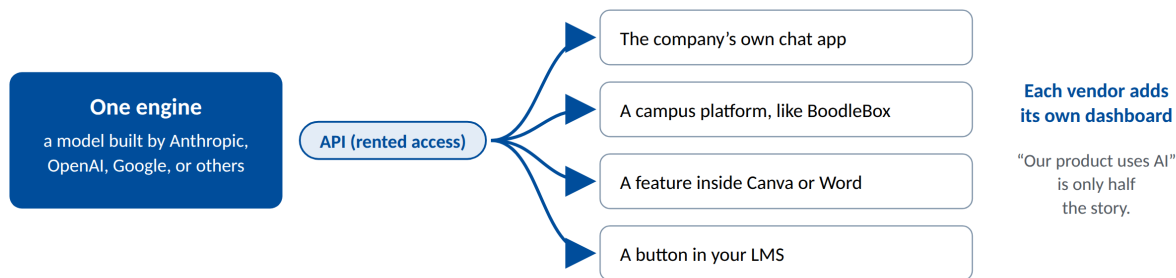
CLARIFICATION: WHAT THE CAR ADDS

In most cases, the interface decides *how* you can work with the engine: whether you can upload a file, save a project, or talk out loud. The model matters, but so do the tools, instructions, and access the app gives it, so the same engine can feel different in a different car. One of those instructions is the **system prompt**, written by the builder before you type anything. You'll probably never see it, but it shapes how the tool behaves.

Demo 1 (3:26): in Claude, the model menu is a drop-down at the bottom right of the chat window, where you can also choose an **Effort** setting (a bit like choosing four-wheel versus two-wheel drive). On a free account, some models or options may be greyed out or missing.

ONE ENGINE, MANY CARS Video 4:35

Companies that build models, like Anthropic, OpenAI, and Google, also rent them to other software companies through an **API** (application programming interface): text goes in, and text or other media comes back. So when a product "uses AI," it very often means the vendor is renting an engine and putting its own dashboard on it.



THREE QUESTIONS FOR ANY “AI-POWERED” PRODUCT

- Which model is under the hood?
- What happens to our data when we use it? (It gets its own module in this series.)
- Who built this engine, and from what?

CLARIFICATION: A CAR CAN CARRY MORE THAN ONE ENGINE

Microsoft 365 Copilot, for example, has run on OpenAI’s models and in 2025 began offering Anthropic’s models as well (Lamanna, 2025). The brand on the button doesn’t always tell you which engine answered.

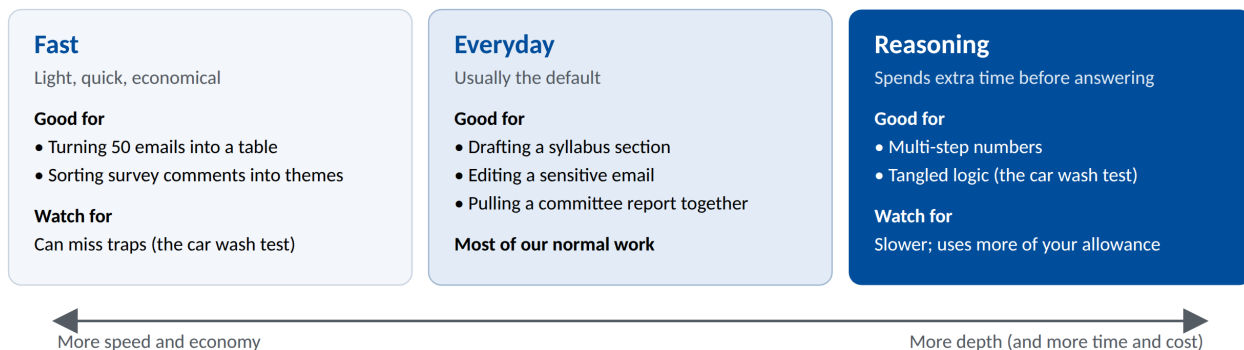
THE FOUR COMPANIES YOU’LL MOSTLY MEET Video 5:41

Most of what you’ll run into comes from **Anthropic** (Claude), **OpenAI** (ChatGPT), **Google** (Gemini), and **Microsoft** (Copilot). The video doesn’t rank them, and most of us would do fine with any of the four. The lead changes hands often, so a few habits help more than any ranking:

- **Pick one and learn it well**, inside and out, and keep one or two others on your radar.
- **Don’t chase the news**. Hopping to whichever tool made headlines this week costs a lot of time and energy.
- **Watch for lock-in**. The longer you work in one vendor’s car, the harder it is to leave. Keep your important work in formats you can take with you.

THREE KINDS OF MODELS: FAST, EVERYDAY, AND REASONING Video 6:55

Inside any one company’s lineup there’s usually a small set of engines that trade speed and economy for depth.



CLARIFICATION: WHAT “THINKING” DOES AND DOESN’T MEAN

On a complicated problem, a reasoning model can work through intermediate results and check parts of its answer, which can improve the result. Some tools show a summary of that work, but the summary isn’t a complete record of what happened, and a longer explanation isn’t proof the answer is right. The categories are shortcuts: training, tools, and the computation used all matter, not just model size.

THE CAR WASH TEST (VIDEO 9:13)

D lives half a mile from the car wash and wants more exercise. To wash the car, should D walk or drive?
Claude’s fast model suggested walking there and driving back, forgetting the car is still in the driveway. The reasoning model said drive, because the car has to be at the car wash. One example isn’t a ranking, but it’s a reason to compare.

Tokens and limits (8:26). A **token** is a small chunk of text the model processes: a whole word, part of a word, or punctuation. As a rough rule for English, one token is about four characters, so 100 tokens is about 75 words (OpenAI, n.d.). Tokens help explain cost and capacity, but your account limit can also depend on the model, tools, attachments, and effort setting. You don’t need to count them; just know that a long chat or a large report can ask the system to do more work.

WHERE YOU USE IT Video 10:03

Place	What to know
Web	Runs in your browser, with nothing to install.
Desktop app	Quicker to reach. On some platforms it can work with files on your computer or see the window you're in, but only if you turn that on.
Phone app	Good for voice, and for snapping a picture of a whiteboard or a mystery object.
Built into your tools	Inside Word, Excel, Canva, or an LMS like Canvas. Often nobody asked for it, and you may not be able to turn it off. Ask which engine is behind the button, and what happens to your text when you press it.

FOUR QUESTIONS FOR CHOOSING Video 11:00

Before all four: What information are you about to share, and is any of it protected or private?

1 How complex?

Sorting and formatting:
a fast model

Most nuanced writing:
an everyday model

2 Multi-step?

Multi-step numbers or
really tangled logic:
**a reasoning model, or
turn up the effort**

Still check the result.

3 How much material?

Know your tool's
context window

Fitting a long report in
doesn't mean it will use
every detail correctly.

4 Where, and what?

Where are you working,
and what can the tool access?

A file on your computer may
still be sent to a cloud
service when AI uses it.

Clarification: context window. This is how much material the model can have available at one time, including room for its response. Working memory is a useful comparison, but it isn't human memory, and a report that fits in the window can still have details used incorrectly. The question that comes before all four, what you're about to share, gets its own module in this series.

CHECK YOURSELF: FIVE TERMS (VIDEO 12:33)

Define each term in a few words, then compare.

Model: the engine that predicts and writes.

Interface: the car around it.

System prompt: the builder's standing instructions.

Token: a chunk of text the model processes, such as a word, part of a word, or punctuation.

Context window: how much material it can have available at one time, including room for its response.

TRY IT THIS WEEK (VIDEO 14:00)

Take one routine task and run it two ways: on a **fast** model and an **everyday** one. Then note three things:

- How long it took
- How much it wrote
- What you'd have to check before using it

Model names will keep changing. The three kinds of models and the four questions are what keep you oriented.

SOURCES

Lamanna, C. (2025, September 24). *Expanding model choice in Microsoft 365 Copilot*. Microsoft. <https://www.microsoft.com/en-us/microsoft-365/blog/2025/09/24/expanding-model-choice-in-microsoft-365-copilot/>

OpenAI. (n.d.). *Understanding and counting tokens*. OpenAI Help Center. Retrieved October 1, 2026, from <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>

This handout was created in collaboration with Claude AI (Anthropic), drawing on the module's final narration. Model names and product features reflect the time of recording and change often.