

Types of AI

Reading time: ~12 min · Prereqs: Lesson 01

Why there are so many different "AIs"

Open any tech site and you'll see ChatGPT next to a spam filter next to a chess engine, all labelled "AI." That's a bit like calling a bicycle, a jumbo jet, and a submarine all "vehicles" — true, but not very useful when you're trying to figure out which one to take to work.

This lesson gives you four simple lenses for slicing up the AI landscape. You don't need to memorise the jargon — just recognise these categories when you see them, and headlines will start making much more sense.

1. By how capable it is: Narrow, General, Super

The most important distinction first, because it's the one the press gets most wrong.

Narrow AI (ANI) is good at *one* kind of task. This is **all AI that currently exists — every single tool you can name today**. A spam filter is brilliant at spotting spam and useless at writing a poem. ChatGPT and Claude handle language tasks; they can't drive your car or diagnose a medical scan on their own. A chess engine beats the world champion — but ask it to play Go and it draws a blank. "Narrow" doesn't mean weak; it means *specialised*.

General AI (AGI) would be a system as flexible as a human — able to learn any task, switch between domains, and reason across them. **AGI does not exist yet**. It is the subject of serious research and enormous controversy. Some researchers think it's decades away; others think it's imminent; others think current definitions make it philosophically impossible to pin down. What you can safely assume: if someone has built AGI, the world will look very different on that day.

Superintelligence (ASI) is a hypothetical system smarter than humans at *everything*. This is science fiction for now. It makes for great movies and urgent blog posts, but

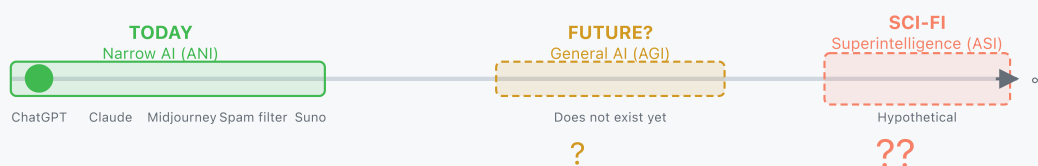
there's nothing to interact with.

The rule for decoding headlines

Whenever a headline says "AI can now [do amazing thing]" — fill in the blank with "narrow AI got better at one specific version of [thing]." "AI beats radiologists at spotting tumours" = a narrow predictive image model trained on thousands of scans performs well on one specific benchmark. That's impressive! And it's not the same as a doctor.

INTERACTIVE: THE CAPABILITY TIMELINE

Hover over each segment to read what's real today vs. hypothetical.



2. By what it produces: Predictive vs Generative

This split cleanly divides the quiet old AI from the loud new AI.

Predictive AI — also called classifying AI — answers "which bucket does this belong in?" or "what number will this be?" Examples: spam-or-not, fraud-or-not, "will this loan default?", weather forecasting, medical image diagnosis, product recommendations. This type of AI has been quietly running inside every major digital service for twenty years. You use it dozens of times a day without noticing.

Generative AI creates new content that didn't exist before: text, images, audio, video, code. ChatGPT and Claude write text. Midjourney and Stable Diffusion paint images.

Suno and Udio compose music. Sora and Runway generate video. GitHub Copilot writes code. This is the category that exploded into public awareness starting in 2022 with DALL-E and ChatGPT. It's newer, it's louder, and it's what most people mean when they say "AI" in casual conversation today.

Both types are narrow

Predictive AI and Generative AI are both Narrow AI. "Generative" doesn't mean smarter or more general — it means the output format is creative content rather than a classification or number. A generative music model can't also write code; a text LLM can't also generate video (unless it's specifically designed to be multimodal).

3. By what kind of content: Modality

"Modality" is just the word for the *type of data* the model works with.

- **Text** — Large Language Models (LLMs): Claude, ChatGPT, Gemini, Llama, Mistral.
- **Image** — generators like Midjourney, DALL-E 3, Stable Diffusion; recognisers like Google Lens.
- **Audio** — speech-to-text tools like Whisper; text-to-speech like ElevenLabs; music generators like Suno and Udio.
- **Video** — generators like Sora, Runway, Pika; analysis tools that can summarise or caption video.
- **Multimodal** — handles multiple types at once. You can send Claude a photo and ask a question about it. GPT-4o can accept text, images, and audio. Most leading models released in 2024–25 are multimodal.

Modality matters when choosing a tool. If you need to transcribe a meeting recording, you want an audio model. If you need to turn a product description into an image, you want an image model. A text-only LLM can't "see" a photo, no matter how good it is at writing.

4. By where it runs: Cloud vs Local

This distinction matters more than most people realise — especially once you think about privacy.

Cloud (closed) models run on a company's servers. You access them over the internet through a browser or app. Claude, ChatGPT, and Gemini all work this way. The upside: these models are powerful (they run on massive clusters of specialised hardware), easy to start (sign up, type, done), and maintained by someone else. The downside: every prompt you send travels over the internet to that company's servers. Most providers have privacy policies that allow them to review conversations for safety, and some use them to improve future models. If you're pasting sensitive documents — contracts, medical notes, confidential strategy — that content is no longer only on your machine.

Local (open) models are downloaded and run entirely on your own computer. *Nothing leaves your machine.* The leading open models include Meta's Llama family, Mistral, Google's Gemma, DeepSeek, and Qwen. They are released under open licences that allow free use. The upside: total privacy, no ongoing cost per query, works offline. The downside: you need hardware. Rough guide: 8 GB RAM runs small models slowly; 16 GB is comfortable for mid-size models; 32 GB+ is needed for the large ones. A dedicated GPU or Apple Silicon Mac makes it much faster.

Two tools make running local models easy enough for non-technical users: **Ollama** (runs in the terminal with one command) and **LM Studio** (a graphical app that feels like ChatGPT, no terminal required). Both are free.

INTERACTIVE: CLOUD VS LOCAL

 **Cloud (Closed)**

 **Local (Open)**



The model runs on the company's servers. You connect over the internet.

✓ Pros

⚠ Cons

- Your prompts leave your machine

- Very powerful — runs on massive hardware
- No setup — sign up and go
- Always up to date
- Multimodal (text + images) out of the box

- Heavy use costs money
- Needs internet connection
- Subject to company's data policies

Example tools: Claude (claude.ai), ChatGPT (chatgpt.com), Gemini (gemini.google.com)

All four lenses in one picture

Any AI tool you encounter can be described on all four axes. Here is the interactive grid — twelve real tools, tagged and filterable.

INTERACTIVE: FILTER REAL AI EXAMPLES

Click filters to highlight matching tools. Multiple filters combine with AND. Use Reset to clear all.

OUTPUT

Generative

Predictive

MODALITY

Text

Image

Audio

Video

Multimodal

WHERE

Cloud

Local

Either

Reset

Chess engine

Predictive

Text

Either

Gmail spam filter

Predictive

Text

Cloud

Claude

Generative

Multimodal

Cloud

Midjourney

Generative

Image

Cloud

Tesla Autopilot

Predictive

Video

Local

Weather forecasting

Predictive

Text

Cloud

ChatGPT

Generative

Multimodal

Suno music

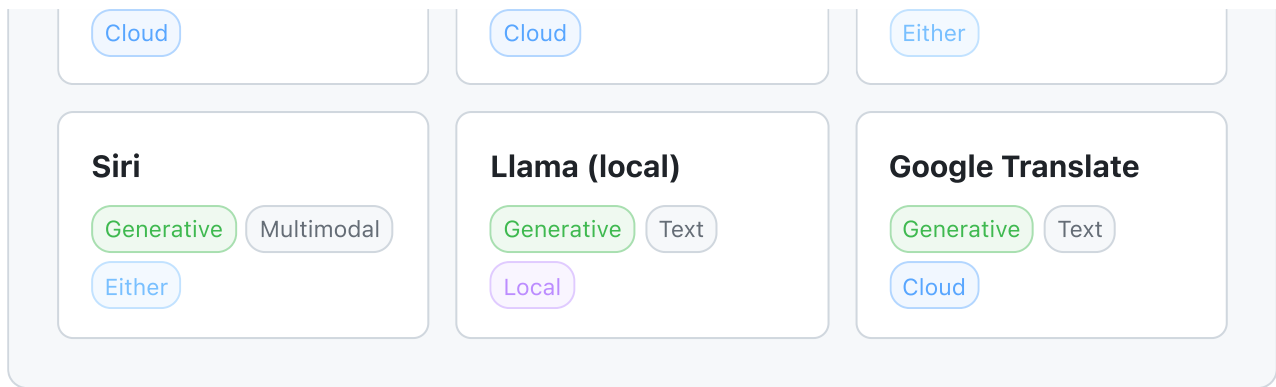
Generative

Audio

Whisper

Predictive

Audio



Putting it together: how to read an AI headline

Here is the four-question checklist to apply to any AI story:

1. **Capability?** Is this real today (narrow), theoretical (AGI), or sci-fi (ASI)?
2. **Output?** Is this creating something new (generative) or sorting/predicting (predictive)?
3. **Modality?** Text, image, audio, video, or several at once?
4. **Where?** Cloud (someone else's servers) or local (your machine, your data)?

Run "AI beats doctors at diagnosing pneumonia from X-rays" through the filter: Narrow (it's a real tool), Predictive (it classifies "pneumonia or not"), Image (it reads X-rays), Cloud (it runs in a hospital IT system). That's a much more concrete, testable claim than just "AI beats doctors."

Which kind should you start with?

If you're brand new to AI: start with a **cloud generative text model**. Claude.ai and ChatGPT are free to start, require no setup, and let you immediately experience the most discussed category of AI today. Once you're comfortable there, the other categories make much more sense in context.

What's coming up

Now you can name the type of AI you're looking at. Next lesson covers the four *doorways* into AI: Chat, API, CLI, and Code tools — same underlying models, four very

different experiences.

LAST-REVIEWED: 2026-06-04