A COMPANION TO UNDERSTAND AI · THE AI MASTERY SERIES

The AI Dictionary, in plain English.

The jargon, actually translated — what each term means, why it matters, and a tiny example where one helps. Nothing to memorise: recognition, not recitation.

Alvoreach · alvoreach.com/glossary · the living version, always current

01 The basics

Artificial intelligence (AI)

Software that does things we used to think needed a human mind — recognising a face, writing a paragraph, answering a question. It is not a mind. It is pattern-matching at enormous scale, and knowing that is the first step to using it well.

Machine learning

The way most modern AI is built. Instead of a person writing every rule, you show the system millions of examples and let it work out the patterns itself. The system is not taught the answer; it is taught to guess, then corrected, millions of times over.

Model

The finished result of that training — the thing you actually use. When people say "ChatGPT" or "Claude", they mean a model: a very large set of learned patterns that turns your input into a plausible output.

> *The model is the engine; the app is the car. One company can offer several models, and one app can let you pick between them.*

Large language model (LLM)

A model trained on a vast amount of text to predict what word comes next. That sounds trivial, but doing it well turns out to require something that looks a lot like understanding language. Every chatbot you have used is an LLM underneath.

Chatbot

The conversation window wrapped around a model. The distinction matters: the chatbot is the product — with its own memory, rules, and tools bolted on — while the model underneath may be the same one powering a dozen other products.

Multimodal

A model that handles more than just text — images, audio, sometimes video — in the same system. Show it a photo and ask a question about it: that is multimodal.

> *Showing a model a photo of your fridge and asking "what can I cook?" works because it is multimodal.*

02 How models work

Neural network

The structure a model is built on: layers of simple calculations connected together, loosely inspired by neurons. No single part is intelligent; the behaviour emerges from billions of small parts adjusted during training.

Token

The small chunk a model actually reads and writes — usually a word or part of a word. Models think in tokens, not letters, which is why they sometimes miscount characters. Pricing and length limits are measured in tokens too.

> *"Unhappiness" might be three tokens: "un", "happi", "ness". When an app mentions "usage limits", it is counting tokens.*

Parameter

One of the many internal dials a model adjusts during training. Modern models have billions of them. More parameters can mean more capability, but the number alone tells you surprisingly little — a headline figure, not a quality score.

> *"A 70-billion-parameter model" just means a big one — and more is not automatically better.*

Weights

The values those dials are actually set to — the learned numbers that make a model that model. Copy the weights and you have copied the model, which is why "open weights" is such a consequential phrase.

Training

The expensive, one-time process of building a model by feeding it data and adjusting it until its guesses get good. It can take months and enormous computing power. Once done, using the model (see inference) is comparatively cheap.

Inference

The act of actually using a trained model to get an answer. Every time you send a prompt and get a reply, that is one inference. It is where the running cost of AI lives.

> *Training built the model last year; every reply you get today is inference.*

Transformer

The architecture nearly every modern language model is built on. Its trick is "attention" — weighing which earlier words matter most for predicting the next one. You do not need the maths to use AI, but this is the engine under the hood.

Context window

How much the model can hold in mind at once — your prompt plus its answer, measured in tokens. Go past it and the earliest text falls out of view. A long conversation that starts forgetting the beginning has run out of context window.

> *Paste a long report and the model can discuss it — until it scrolls off the window and is forgotten. "That fell out of the context window" is a far more precise complaint than "it forgot".*

Temperature

A setting that controls how adventurous a model's answers are. Low temperature: focused and repetitive. High temperature: varied and creative, but more likely to wander. It is a dial between reliable and surprising.

Embedding

A way of turning text into a list of numbers that captures its meaning, so a computer can measure how similar two pieces of text are. It is the quiet machinery behind search, recommendations, and RAG.

---

03 Using models well

Prompt

What you type in. The instruction, question, or context you give the model. Most of the difference between a useless answer and a great one is in the prompt — which is exactly why the "why it works" matters more than any prompt you copy.

> *"Write a thank-you note to a colleague" is a prompt. Phrasing better ones is a skill in its own right.*

System prompt

A hidden instruction that shapes how a model behaves across a whole conversation — its role, tone, and rules — set before you type anything. It is the difference between the same model acting as a terse coder or a patient tutor.

> *"You are a helpful, concise assistant", set behind the scenes, is a system prompt; your typed request is just the prompt.*

Prompt engineering

The craft of writing prompts that work reliably — structure, examples, constraints, and knowing what the model needs to be told versus what it can infer. Less magic than the name suggests; closer to writing a good brief for a capable colleague.

Zero-shot & few-shot

Asking a model to do a task cold (zero-shot) versus showing it a few worked examples first (few-shot). The difference is often dramatic — a handful of good examples can outperform a page of instructions.

RAG (retrieval-augmented generation)

Giving a model access to a specific set of documents and letting it look things up before answering, instead of relying only on what it learned in training. This is how you get an AI that can answer questions about your own files without retraining it.

Fine-tuning

Taking an already-trained model and training it a little further on a narrower set of examples, to specialise it — a general model nudged toward legal writing, say, or your company's tone. Cheaper than starting over, and often enough.

> *Fine-tune a general model on a law firm's past contracts and it drafts in that firm's style.*

Agent

An AI set up not just to answer but to act — to use tools, take steps, and work toward a goal with less hand-holding. Powerful and worth understanding precisely because "less hand-holding" is also where the risk moves.

> *Instead of telling you how to book a table, an agent goes through the steps to book it.*

04 The jagged edges

Hallucination

When a model states something false with complete confidence. It is not lying — it has no concept of true or false. It is producing the most plausible-sounding continuation, and sometimes the most plausible thing is wrong. This is the jagged edge to watch.

> *The classic case is an invented citation that looks perfectly real — the danger is that there is no wobble in the tone to warn you.*

Bias

Models learn from human-made data, so they inherit human patterns — including the unfair ones. Bias in AI is not a bug that got in; it is the training data faithfully reflected. The fix starts with knowing it is there.

Guardrails

The rules and filters wrapped around a model to stop harmful, off-limits, or off-brand output. Necessary, imperfect, and worth understanding: they shape what a model will and won't say far more than most users realise.

Alignment

The work of making a model actually pursue what its makers and users intend — helpful, honest, harmless — rather than what its raw training would produce. The gap between "capable" and "trustworthy" is the alignment problem.

05 The ecosystem

API

The plug socket that lets other software use a model directly — no chat window involved. When an app "has AI", it is usually calling a model through an API. Metered and billed by the token.

Open weights

A model whose trained parameters are published, so anyone can download and run it themselves rather than renting it through an API. Often loosely called "open source". It changes who controls the model — and who is responsible for it. Open versus closed is really a privacy-and-control question: does your data leave your machine, and who depends on whom?

> *A closed model is like a streaming service you log into; an open model is like a file you own and can run on your own computer.*

GPU & compute

The specialised chips (GPUs) and raw processing power (compute) that AI runs on. Compute is the scarce resource behind the headlines — who has it, who needs it, and what it costs decides much of the industry.

Benchmark

A standard test used to compare models — maths problems, coding tasks, exam questions. Useful, and gameable: a high score means the model is good at the benchmark, which is not always the same as good at your job.

AGI (artificial general intelligence)

The hypothetical model that matches humans across the board rather than at specific tasks. Nobody agrees on the definition, the date, or whether the term is even useful — treat every confident AGI claim, for or against, with the same care.

New terms will keep arriving.

That is permanent in this field. The real skill is not knowing them all, but the two-minute habit of asking for a plain definition with a tiny example whenever one stops you. The living version of this dictionary grows at **alvoreach.com/glossary** — and the whole story is in *Understand AI*, whose first chapter is free at **alvoreach.com/start**.