



Nocturne

The 10 AI words everyone uses and nobody explains

A plain-English glossary. No jargon, no maths, no assumed knowledge.

Ten minutes and you'll follow any AI conversation in the room.

Model

THEY SAY

"They released a new model." · "Which model are you using?"

IT MEANS

The thing that does the thinking. When you use ChatGPT or Claude, the app is the window and the model is what's behind it — the part that takes what you typed and produces an answer.

PICTURE IT

A model is an employee, and the app is their desk. Companies keep hiring smarter employees; the desk looks the same. That's why the app you use barely changes but people keep saying it "got better".

WHY YOU CARE

Almost every AI headline is about a new model. When someone names one, they're naming an employee, not a product. Different models are better at different jobs, and picking the right one is most of what "using AI well" means in practice.

Token

THEY SAY

"That'll cost you a lot of tokens." · "It ran out of tokens."

IT MEANS

A chunk of text, usually about three-quarters of a word. AI systems don't read letters or words — they read tokens. "Unbelievable" might be three tokens; "cat" is one.

PICTURE IT

Tokens are the AI's units the way minutes are a phone plan's units. Nobody bills you per conversation; they bill you per minute. Same here — you're billed per chunk of text going in and coming out.

WHY YOU CARE

It's the meter. Anything you're charged for is measured in tokens, and it counts both what you send and what you get back. A long document pasted into a chat costs real money before the AI has said a word.

Context window

THEY SAY

"It has a million-token context window."

IT MEANS

How much the model can hold in mind at once — your current conversation, any documents you've pasted, and its own replies so far. Once it's full, the oldest things fall out.

PICTURE IT

A desk, not a filing cabinet. Everything spread out on the desk is available instantly. When the desk is full and you add something new, something slides off the far edge and is simply gone — and the model doesn't know it forgot.

WHY YOU CARE

This explains the most common frustration people have with AI: *why has it forgotten what I told it?* It hasn't got worse. The desk filled up. Long conversations get less reliable at the start, not the end.

Prompt

THEY SAY

"That's a prompting problem." · "Prompt engineering."

IT MEANS

What you type. That's it. The whole instruction — your question, your context, your examples, your "be concise" at the end.

PICTURE IT

A brief to a very fast, very well-read contractor who has never met you, knows nothing about your business, and will not ask clarifying questions unless invited. Vague brief, generic work.

WHY YOU CARE

The gap between people who find AI useless and people who find it indispensable is mostly here, and it isn't a technical skill. It's the ordinary skill of explaining what you want — what the output is for, who reads it, what "good" looks like, and one example.

05 • the important one

Hallucination

THEY SAY

"It hallucinated the whole thing."

IT MEANS

When the model states something false with total confidence. Invented sources, made-up quotes, a plausible-sounding fact that isn't.

PICTURE IT

Someone who would rather guess smoothly than admit they don't know. Not lying — lying needs awareness of the truth. It's closer to someone finishing your sentence for you, and being wrong.

WHY YOU CARE

This is the most important thing in this document. AI does not have a "not sure" face. Wrong answers arrive with exactly the same confidence as right ones. The practical rule: **anything you'd be embarrassed to be wrong about, check.** Names, numbers, dates, quotes, legal or medical claims, and anything with a link.

Inference

THEY SAY

"Inference costs are falling."

IT MEANS

The model actually running — the moment it takes your question and produces an answer. Distinct from training, which is the months of preparation that happened long before you showed up.

PICTURE IT

Training is the years someone spent at medical school. Inference is the ten-minute appointment. The training was ruinously expensive and happened once; the appointment is cheap and happens millions of times.

WHY YOU CARE

When headlines say AI is "getting cheaper", this is almost always what's got cheaper — and it has fallen dramatically. It's why tools that once gave you a handful of queries for £20 a month now give you hundreds.

Weights

THEY SAY

"They released the weights." · "Open-weight model."

IT MEANS

The model's learned settings — billions of numbers, tuned during training, that hold everything it knows. The weights are the model; everything else is packaging.

PICTURE IT

A recipe developed over years of tasting and adjusting. The weights are the final quantities. Publish them and anyone can cook the dish in their own kitchen; keep them secret and people can only eat at your restaurant.

WHY YOU CARE

It's what "open source AI" actually turns on. When a company releases the weights, anyone can run that model themselves, free, forever, with no company able to withdraw it or change the price. That's why those releases are treated as a big deal.

Fine-tuning

THEY SAY

"We *fine-tuned* it on our data."

IT MEANS

Taking a finished, general-purpose model and training it a bit further on a narrower set of examples, so it gets better at one specific job or adopts a particular style.

PICTURE IT

Hiring an experienced professional and then spending two weeks teaching them your company's way of doing things. You're not teaching them their profession — you're teaching them your house style.

WHY YOU CARE

Mostly so you can recognise when you don't need it. It's expensive and slow, and people reach for it when the real answer is a better prompt or the next term in this list. It genuinely helps with consistent tone and format. It does not reliably teach a model new facts.

RAG

retrieval-augmented generation

THEY SAY

"We put RAG on top of it."

IT MEANS

Letting the model look things up before answering. Your documents get stored somewhere searchable; when you ask a question, the relevant few pages are fetched and handed to the model along with your question.

PICTURE IT

An open-book exam. The model isn't remembering your company handbook — someone is sliding the right page across the desk at the right moment.

WHY YOU CARE

This is how "chat with your own documents" works, and it's the honest answer to most business AI questions. Cheaper than fine-tuning, updates the moment you update a document, and it can show which page an answer came from — the only practical defence against term 05.

Agent

THEY SAY

"Agentic AI is the next big thing."

IT MEANS

An AI that does a task in several steps rather than producing one reply. It can use tools — search the web, read files, run software, send an email — decide what to do next based on what it found, and keep going until the job is done.

PICTURE IT

The difference between asking someone for advice and asking them to handle it. Advice comes back in one message. Handling it means they go away, make phone calls, hit a snag, adjust, and return when it's finished.

WHY YOU CARE

It's the whole current wave, and it changes the risk. A chatbot that's wrong produces a bad paragraph you can ignore. An agent that's wrong takes wrong actions — sends the email, edits the file, spends the money. That's why agents come with permission prompts, and why clicking "allow" without reading is the modern equivalent of signing without looking.

That's the vocabulary.

Ten words is genuinely most of it. The rest of AI is people using these ten in combinations to sound more certain than they are.

Every Tuesday I send one email: what actually happened in AI that week, in this same plain English, plus one thing you can use. Three minutes.

You're already subscribed — the first one lands Tuesday.

— Noc

Reply to it any time. I read them all.