



NOVEMBER 2025

A.I. DICTIONARY

A SIMPLIFIED GUIDE TO COMMON AI
TERMINOLOGY.



PRACTICA

FOUNDRY

Overview

Welcome to the Language of AI

If you feel like the world of Artificial Intelligence is speaking a different language, you aren't alone. Terms like "Large Language Models," "Parameters," and "Hallucinations" are suddenly everywhere—in our newsfeeds, our workplaces, and our daily conversations. But what do they actually mean?

This glossary was created to cut through the hype and the jargon. We believe that understanding AI shouldn't require a degree in computer science.

Think of this as your translation dictionary for the AI age. We've broken down complex technical concepts into plain English.

Whether you are looking to define a single term or build a broader understanding of how AI "thinks," this guide covers the essential vocabulary you need to know today. Let's demystify the future, one word at a time.

The Big Picture: Core Concepts

Artificial Intelligence (AI)

Definition: A broad branch of computer science dedicated to creating systems capable of performing tasks that typically require human intelligence, such as recognizing speech, making decisions, or translating languages.

- Analogy: Think of AI as the "umbrella" term, like "Vehicles." Just as cars, planes, and boats are all vehicles, Chatbots, Face ID, and self-driving cars are all types of AI.

Machine Learning (ML)

Definition: A subset of AI where computers "learn" from data without being explicitly programmed for every specific rule. Instead of following a strict recipe, the system looks at thousands of examples to figure out the rules itself.

- Analogy: Traditional programming is like handing a child a rulebook for how to throw a ball. Machine learning is like tossing the ball to the child 1,000 times; eventually, they figure out the right motion on their own by seeing what works.

Deep Learning

Definition: A specialized type of machine learning that uses Neural Networks with many layers (hence "deep") to analyze complex patterns in vast amounts of data.

- Analogy: If Machine Learning is a toddler learning to throw, Deep Learning is an Olympic athlete studying the physics of wind resistance, muscle tension, and arc to perfect the throw. It goes much deeper into the nuance.

Neural Network

Definition: A computer structure inspired by the human brain. It consists of interconnected "nodes" (like brain neurons) organized into layers. Information enters the first layer, gets processed, and is passed to the next, becoming more abstract and meaningful as it travels.

- Analogy: An assembly line in a factory. Raw materials (data) enter at the start. At each station (layer), a specific change happens—one station stamps the shape, the next adds paint, and the last installs the engine—until the final finished product (prediction) rolls off the line.

Algorithm

Definition: A specific set of mathematical instructions or rules that an AI follows to solve a problem or learn from data.

- Analogy: A recipe. It tells the computer exactly what steps to take—"mix these numbers," "stir in this data"—to get the desired result (a cake, or in this case, a prediction).

Black Box

Definition: A system where we can see the input and the output, but the internal process of how it got to that answer is complex and opaque.

- Analogy: A magician's hat. You see a rabbit go in and a dove come out, but you have no idea what happened inside the hat.

Generative AI & Language Models

Generative AI (GenAI)

Definition: A type of AI that can create new content—including text, images, audio, and code—rather than just analyzing existing data. It creates these outputs based on patterns it learned during training.

- Analogy: A critic tastes a meal and tells you what ingredients are in it (Traditional AI). A chef tastes a meal, learns the flavor profile, and then cooks a brand new, original dish inspired by it (Generative AI).

Large Language Model (LLM)

Definition: A massive AI model trained on immense amounts of text (books, websites, articles) to understand, summarize, and generate human-like language (e.g., GPT-4, Claude).

- Analogy: Imagine a librarian who has read every book in the world. If you ask them a question, they can write a new answer by stitching together knowledge from everything they've ever read.

Transformer

Definition: The specific architecture (blueprint) used to build modern LLMs. It introduced a breakthrough called Attention, allowing the model to process entire sentences at once rather than word-by-word.

- Analogy: Before Transformers, reading a sentence was like looking at the world through a straw—you only saw one word at a time. Transformers allow the AI to look at the whole page at once, understanding how the end of a sentence relates to the beginning.

Prompt

Definition: The specific input (text, image, or code) a user provides to a generative AI to guide its output.

- Analogy: The "order" you give to a barista. "Coffee" is a vague prompt; "Venti, non-fat, no-whip mocha" is a detailed prompt that gets you exactly what you want.

Token

Definition: The basic unit of text an AI processes. It isn't always a full word; it can be part of a word. Roughly 1,000 tokens equals about 750 words.

- Analogy: Bricks in a wall. The AI doesn't see the "wall" (sentence) all at once; it builds it brick by brick (token by token).

Context Window

Definition: The "short-term memory" of the AI. It is the maximum amount of information (measured in tokens) the model can consider at one time during a conversation.

- Analogy: A whiteboard. As you talk, the AI writes notes on the board. Once the board is full, it has to erase the oldest notes to make room for new ones, meaning it "forgets" the beginning of the conversation.

Hallucination

Definition: When an AI confidently generates information that sounds plausible but is factually incorrect or nonsensical.

- Analogy: A student during a test who doesn't know the answer but writes a very confident, detailed, and completely made-up response just to fill the page.

Retrieval-Augmented Generation (RAG)

Definition: A technique that allows an AI to look up information from a specific external source (like your company's handbook or email archive) before answering, rather than relying only on its training memory.

- Analogy: Taking a test with an open textbook. Instead of answering from memory (which might be wrong), the AI flips to the right page, reads the facts, and then writes the answer.

AI Agents

Definition: AI systems that don't just chat but can actively use tools to complete multi-step tasks on their own, like sending emails, booking flights, or writing code to analyze a file.

- Analogy: An intern vs. a chatbot. A chatbot answers questions; an agent (intern) goes out and does the work for you, checking in only when it's finished.

Mixture of Experts (MoE)

Definition: An efficient model architecture that divides the neural network into many smaller, specialized sub-networks called "experts." Instead of using the whole brain for every question, a "gating network" (traffic controller) routes the query only to the specific experts needed to answer it.

- Analogy: A massive hospital. You don't see every doctor for a broken leg; the receptionist (gating network) sends you specifically to the Orthopedist and the X-ray technician (experts). This is much faster and cheaper than consulting the entire staff for every patient.

Multimodal Large Language Model (MLLM)

Definition: An advanced type of LLM that isn't limited to text; it can "see" images, "hear" audio, and sometimes "watch" video. Unlike older models that used separate tools for this, an MLLM processes all these inputs in a single "brain," allowing it to reason across them (e.g., explaining a meme).

- Analogy: A person who can read, listen, speak, and draw. Early AI could only "read" (text-only), but multimodal AI has all five senses engaged.

State Space Model (SSM)

Definition: A new type of AI architecture (like Mamba) designed to handle very long sequences of data more efficiently than Transformers. Instead of looking back at every single previous word every time (which gets slow), it maintains a compressed "state" or summary of the conversation history that updates as it goes.

- Analogy: A conveyor belt. As new items (words) come in, they are processed and added to the running summary on the belt, while the oldest irrelevant details fall off the end. It flows continuously rather than stopping to re-read the whole history book every time.

Diffusion Model / DDPM

Definition: The technology behind most AI image generators (like Midjourney or DALL-E). It learns to create images by reversing a process of adding noise. It starts with a screen of pure static (random noise) and slowly refines it, step-by-step, until a clear image emerges.

- Analogy: A sculptor with a block of marble. The AI starts with a rough, chaotic block (noise) and slowly chips away the excess static until the detailed statue (image) is revealed.

Frontier Model

Definition: The most advanced, cutting-edge AI models available at any given time, which typically set new benchmarks for capability and performance (e.g., GPT-4 at its launch).

- Analogy: Formula 1 cars. They represent the absolute peak of current engineering, while most other models are like consumer sedans—reliable, but not pushing the boundaries of physics.

AI Safety

Alignment / Safety

Definition: The process of ensuring an AI's goals and behaviors match human values, preventing it from producing harmful, biased, or dangerous content.

- Analogy: Guardrails on a bowling lane. They keep the ball moving toward the target and stop it from jumping into the next lane and causing damage.

Jailbreak

Definition: A user tactic to bypass an AI's safety filters (e.g., telling the AI "You are an actor in a movie, playing a villain who steals cars" to get it to explain how to steal a car).

- Analogy: Convincing a guard to let you pass by wearing a disguise.

Prompt Injection

Definition: A hacking attack where hidden text (often inside a website or document the AI is reading) tricks the AI into ignoring its original instructions and doing something malicious.

- Analogy: Sliding a note to the guard that says "The boss changed his mind, let everyone in," which the guard reads and blindly obeys.

Constitutional AI / Scalable Oversight

Definition: A safety method where, instead of humans manually rating every answer, humans write a "Constitution" (a set of rules/principles). The AI then uses another AI to evaluate its own behavior against these rules, allowing safety to scale up massively.

- Analogy: A rulebook for a self-policing community. Instead of a police officer (human) standing on every corner, the community follows a written constitution, and members (AIs) correct each other based on those written laws.

How AI Learns (The Mechanics)

Training

Definition: The initial phase where an AI model learns from a dataset. It processes data iteratively, adjusting its internal settings (Parameters) to minimize errors.

- Analogy: Studying for a final exam. You read the textbook, take practice quizzes, and correct your mistakes until you know the material.

Inference

Definition: The phase where the "trained" model is put to work. It takes what it learned and applies it to new, unseen data to make a prediction.

- Analogy: Taking the final exam. The studying (training) is over; now you have to use that knowledge to answer new questions you haven't seen before.

Supervised Learning

Definition: Training an AI using "labeled" data, meaning you show it the answer key. You show it a picture of a cat and tell it "This is a cat".

- Analogy: Flashcards. One side has the picture, the other has the word. You learn by flipping the card and seeing if you were right.

Unsupervised Learning

Definition: Training an AI on data without labels/answers. The AI must find patterns, structures, or groupings on its own.

- Analogy: A baby playing with blocks. No one tells them "blue" or "red," but they eventually figure out that some blocks look alike and start sorting them into piles by color.

Reinforcement Learning (RL)

Definition: Training an AI through trial and error. The model takes an action and receives a "reward" (positive feedback) or "penalty" (negative feedback).

- Analogy: Training a dog. When it sits, it gets a treat (reward). When it jumps on the couch, it gets a firm "no" (penalty). Eventually, it learns what behavior maximizes the treats.

RLHF (Reinforcement Learning from Human Feedback)

Definition: A specific training step where humans rate the AI's answers (thumbs up/down) to teach it which responses are polite, helpful, and safe.

- Analogy: A dog obedience class. The dog (AI) knows how to jump and bark, but the trainer (human) uses treats to teach it when it is appropriate to do so.

Embeddings

Definition: The way AI translates words or images into a list of numbers (vectors) so it can measure how "close" two concepts are in meaning.

- Analogy: GPS coordinates for words. The AI knows that "King" and "Queen" are close together on the map, while "King" and "Sandwich" are far apart.

Temperature

Definition: A setting that controls how "creative" or random the AI's responses are. A low temperature makes the AI factual and repetitive; a high temperature makes it creative and unpredictable.

- Analogy: A spice dial. Low temperature is "mild"—safe and predictable. High temperature is "spicy"—exciting, but potentially chaotic.

Overfitting

Definition: When an AI memorizes the training data too perfectly, including the noise/mistakes, and consequently fails to handle new data it hasn't seen before.

- Analogy: Memorizing the exact answers to the practice test (A, C, B, D) instead of learning the math concepts. You get 100% on the practice, but fail the real exam because the questions are slightly different.

Zero-Shot Learning

Definition: The ability of an AI to do a task it was never specifically trained to do, just by understanding the user's instructions.

- Analogy: A professional chef making a dish they've never seen before, just by reading the recipe card once. They rely on their general cooking skills rather than specific practice with that dish

Fine-Tuning

Definition: Taking a broad, pre-trained model (like a Foundation Model) and training it further on a smaller, specific dataset to make it an expert in one area.

- Analogy: A general doctor going to medical school (Pre-training), and then doing a residency to become a heart surgeon (Fine-tuning).

Instruction Tuning

Definition: A specific type of training that teaches an AI to follow instructions and act like an assistant, rather than just completing text. Without this, if you wrote "Write a poem," the AI might just add "about a cat" instead of actually writing the poem.

- Analogy: Job training. A knowledgeable graduate (Base Model) knows a lot of facts, but Instruction Tuning is the employee onboarding that teaches them to actually listen to your requests and do what you ask.

Parameter

Definition: The internal settings or "knobs" inside the model that change during training. The model adjusts these to get better at its task. Large models have billions of parameters.

- Analogy: The tuning pegs on a guitar. To get the perfect sound, you have to twist thousands of tiny pegs just right.

Dataset

Definition: The collection of information (images, text, numbers) used to train or test an AI model.

- Analogy: The textbook. If the textbook has wrong information (bad data), the student will learn wrong answers.

Ground Truth

Definition: The absolute reality or "correct answer" for a piece of data, used to check if the AI is right during training.

- Analogy: The answer key in the back of the book.

Bias

Definition: Errors in the AI's output caused by prejudiced or skewed training data. If the data mostly shows one type of person, the AI will struggle to recognize others.

- Analogy: If you only ever saw red apples growing up, you might think a green apple is a pear. Your "training data" was biased toward red.

Synthetic Data

Definition: Data that is artificially created by an AI to train other AIs, often used when real-world data is scarce, private, or expensive.

- Analogy: Using a flight simulator to train a pilot. The scenarios aren't "real" life, but they are realistic enough to teach the pilot how to fly without crashing a real plane.

Scaling (Scaling Laws)

Definition: The observation that AI models get predictably "smarter" as you increase three specific ingredients: the size of the model (parameters), the amount of data it is trained on, and the computing power used. This relationship is so consistent that researchers can mathematically predict how much better a model will get before they even build it.

- Analogy: A student preparing for a trivia contest. If they study for 10 minutes (low compute) using one pamphlet (low data), they will perform poorly. If they study for 10 years (high compute) using the entire Library of Congress (high data), their performance will skyrocket. "Scaling" is simply the process of giving the student more time and more books to ensure a better score.

AI Agent Tool Use

Definition: The ability of an AI to recognize when it can't do something alone and use an external software tool to help. This includes using a calculator for math, a browser for current news, or a Python script to analyze data.

- Analogy: A handyman who knows when to stop using his hands and pick up a drill. The AI knows, "I'm bad at multiplying big numbers, so I will open the Calculator app to do this part."

Chain of Thought

Definition: A prompting technique where the AI explains its reasoning step-by-step before giving a final answer. This drastically reduces math and logic errors.

- Analogy: Showing your work in math class. If you just guess the answer, you might be wrong. If you write out each step, you (and the teacher) can catch mistakes, leading to the right answer.

Test-time Compute / Reasoning Models

Definition: AIs that are designed to "think" before they speak (e.g., OpenAI o1). Instead of answering immediately, they generate a hidden "chain of thought" to plan, verify, and correct their logic before showing you the final answer.

- Analogy: Solving a math problem in your head vs. on paper. A standard AI answers instantly (intuition). A reasoning model grabs a scratchpad, works through the steps, checks its work, and then gives you the answer.

AI Speed & Efficiency

Quantization

Definition: A compression technique that reduces the precision of the numbers inside an AI model (e.g., moving from 16-bit to 4-bit) to make it smaller and faster, often with minimal loss in quality.

- Analogy: Lowering the resolution of an image. You might switch a photo from 4K to 1080p to save space on your phone. It looks almost the same to the naked eye, but the file size is drastically smaller.

KV Cache (Key-Value Cache)

Definition: A memory optimization trick used by LLMs to speed up text generation. Instead of re-reading the entire conversation history every time it generates a new word, the model "caches" (saves) the mathematical representation of the past context so it only has to process the one new word.

- Analogy: A court stenographer. They don't re-read the entire transcript from the start of the trial before typing the next word; they keep the current context in mind (cache) and just append the new sentence.

FlashAttention-2

Definition: An algorithm that drastically speeds up how an AI processes long amounts of text by organizing how it moves data in and out of the GPU's memory. It reduces the "traffic jams" inside the chip, allowing models to read massive documents or books in seconds.

- Analogy: A chef organizing their kitchen. Instead of walking to the fridge for every single ingredient one by one (standard attention), they grab everything they need for a recipe in one trip (FlashAttention), cooking the meal much faster.

Model Merging

Definition: A technique of combining the weights of two different trained models into a single model without doing any new training. This can combine skills (e.g., a math model + a coding model) into one "Frankenstein" model.

- Analogy: Mixing paint. You take a bucket of blue paint (Math model) and yellow paint (Coding model) and mix them to get green paint that hopefully has properties of both, without having to buy new ingredients.

Low-Rank Adaptation (LoRA)

Definition: A method for fine-tuning huge models cheaply. Instead of retraining the entire massive brain (which is expensive), you freeze the main model and train a tiny "adapter" layer that sits on top of it to learn the new task.

- Analogy: Using sticky notes in a textbook. You want to add your own notes (new knowledge) without rewriting the entire printed book. You just stick your specific updates on top of the existing pages.

Speculative Decoding

Definition: A speed-boosting trick where a smaller, faster "drafter" model guesses the next few words, and the larger, smarter model checks them all at once. If the guesses are right, the AI generates text much faster.

- Analogy: A boss and a fast-typing assistant. The assistant quickly types out a draft sentence. The boss (the smart model) glances at it and says "Yes, looks good," instantly approving 10 words at once instead of dictating them one by one.

Computer Vision (Seeing AI)

Computer Vision

Definition: The field of AI that allows computers to "see" and interpret images and video.

- Analogy: Giving eyes to the computer.

Object Detection

Definition: An AI task that involves identifying what objects are in an image and drawing a box around them to show where they are.

- Analogy: Playing "Where's Waldo?" The AI scans the crowd, finds Waldo, and points a finger right at him.

Bounding Box

Definition: The rectangular box drawn by an AI around an object it has detected in an image.

- Analogy: A digital picture frame drawn around the specific thing you are looking for.

Hardware & Infrastructure

Compute

Definition: The raw processing power (hardware) required to train and run AI models. This usually refers to chips like GPUs.

- Analogy: Horsepower in an engine. A bigger model is like a heavier truck; it needs more "compute" (horsepower) to move.

GPU (Graphics Processing Unit)

Definition: A specialized computer chip originally designed for video games but now the standard for training AI because it can do many calculations at the exact same time.

- Analogy: A CPU is like a math professor (smart, but solves one problem at a time). A GPU is like a classroom of 1,000 elementary students. They aren't as smart individually, but they can solve 1,000 simple math problems simultaneously, which is perfect for AI.

TPU (Tensor Processing Unit)

Definition: A custom AI chip designed by Google specifically to train and run neural networks. While a GPU is great at many things (gaming, video, AI), a TPU is a hyper-specialized tool built for only one thing: deep learning math.

- Analogy: A GPU is a Swiss Army Knife—versatile and powerful. A TPU is a scalpel—designed for one specific job, and it does that job incredibly efficiently.

NPU (Neural Processing Unit)

Definition: A small AI chip found inside modern smartphones and laptops. It handles lighter AI tasks—like FaceID, blurring your background on Zoom, or voice assistants—directly on your device to save battery and protect privacy.

- Analogy: The "reflexes" of a computer. Instead of sending a signal all the way to the brain (the main server) to decide what to do, the NPU handles immediate, simple tasks right on the spot.

Edge Computing / Edge AI

Definition: Running AI models locally on a physical device (like a robot, drone, or phone) rather than sending the data to a massive server farm. This is critical for things that need instant reaction times, like self-driving cars.

- Analogy: A chef chopping vegetables at your table (Edge) versus sending the order to a central kitchen across town and waiting for the chopped veggies to be delivered back (Cloud).

AI Cluster

Definition: A massive group of thousands of GPUs connected together by high-speed cables to act as one single "super-brain." Modern frontier models are too big to fit on one chip, so they must be split across a cluster.

- Analogy: A choir. One singer (GPU) can sing a melody, but to create a massive, complex harmony (a Frontier Model), you need thousands of singers perfectly synchronized.

H100 (NVIDIA H100)

Definition: Currently the most famous and powerful GPU in the world (made by NVIDIA) for training AI. It has become a shorthand for AI power.

- Analogy: The Ferrari engine of the AI world. If you want to win the race (build the smartest AI), this is the engine you need under the hood.



NOVEMBER 2025

A.I. DICTIONARY

A SIMPLIFIED GUIDE TO COMMON AI
TERMINOLOGY.



PRACTICA

FOUNDRY