UNDERSTAND AI THROUGH MOVIES

100 AI terms explained with films and shows you already know.

One plain definition, one familiar scene and, where it matters, the point the analogy can't carry. Plus a 48-term quick reference for building with AI and using it in the clinic.



WHY FILMS?

AI comes with its own vocabulary, and it grows every month. Most definitions are accurate but hard to remember: they explain one new word with five more.

A good scene is easy to remember. If you have seen Leonard in *Memento* working only from the notes in front of him, you already understand why an AI loses track of a long conversation. The word “context window” then has something to hold on to.

So each entry gives you three things: what the term means, a scene that makes it stick, and, where it matters, the point where the comparison stops working.

HOW TO READ AN ENTRY

019

CONTEXT WINDOW

Memento

The maximum amount of text, counted in tokens, a model can work with in one request.

> THE SCENE
>
> Leonard can only use what is in front of him now: his notes, photos and tattoos. Everything else is gone.

KEEP IN MIND Information outside the window does not exist for the model.

Definition. The plain meaning. Read this first.

The scene. The picture that helps you remember it.

Keep in mind. A practical point, or where the analogy breaks.

A NOTE ON THE ANALOGIES

Every comparison here is a memory aid, not a technical description. Trust the definition when the two disagree. Film titles are used only as references; this guide is not connected with any studio. Definitions follow common usage in 2026; check **developers.google.com/machine-learning/glossary** for formal ones.

CONTENTS

01	AI fundamentals	001-010	04
02	How language models work	011-020	06
03	How models learn	021-030	08
04	Cheaper and faster	031-040	10
05	Prompting and reasoning	041-050	12
06	Memory and retrieval	051-060	14
07	AI agents	061-070	16
08	Testing and failure	071-080	18
09	Images, voice and video	081-090	20
10	The big questions	091-100	22
+	Quick reference: 48 more terms		24

SHORT ON TIME? START WITH THESE EIGHT

They explain most of what confuses people about AI tools today.

MEMENTO
CONTEXT WINDOW
Why AI “forgets”

ROCKY
TRAINING VS INFERENCE
Learning vs using

THE USUAL SUSPECTS
HALLUCINATION
Why fluent can be false

THE DA VINCI CODE
RAG
How AI uses your documents

MISSION: IMPOSSIBLE
AI AGENT
From answers to actions

INCEPTION
PROMPT INJECTION
The main agent security risk

GROUNDHOG DAY
OVERFITTING
Why test scores mislead

EDGE OF TOMORROW
REINFORCEMENT LEARNING
Learning by trial and error

01 AI FUNDAMENTALS

The ten words every other term is built on.

001

ARTIFICIAL INTELLIGENCE

Iron Man

Software that does tasks we usually link to human thinking: understanding language, recognising images, making decisions.

> THE SCENE
> JARVIS hears what Tony asks, works out what he means and acts on it.

KEEP IN MIND Most real AI does one narrow job well. JARVIS does everything.

002

MACHINE LEARNING

The Karate Kid

Software that learns patterns from examples instead of following rules a programmer wrote by hand.

> THE SCENE
> Mr Miyagi never hands Daniel a rulebook. Daniel learns from thousands of repetitions.

003

DEEP LEARNING

Charlie and the Chocolate Factory

Machine learning with many layers in sequence. Each layer turns its input into a slightly more useful form.

> THE SCENE
> Ingredients move from room to room. Each room changes them a little until a finished bar comes out.

004

NEURAL NETWORK

Inside Out

A web of simple connected units. Each unit combines weighted signals from other units and passes a result on.

> THE SCENE
> Several voices push at Riley's control console at once. What she does depends on how hard each one pushes.

KEEP IN MIND Real units are just numbers. They have no personalities.

005

LARGE LANGUAGE MODEL (LLM)

Arrival

A neural network trained on huge amounts of text to predict and generate language.

> THE SCENE
> Louise learns an alien language by studying its patterns across many examples, until she can read and write it.

006

TOKEN

The Lego Movie

The small piece of text a model reads and writes: a word, part of a word, or a punctuation mark.

> THE SCENE
> Everything in Bricksburg is built from bricks. Everything a model reads is built from tokens.

007

TOKENIZATION

The Lego Movie

The step that splits text into tokens and gives each one an ID number.

> THE SCENE
> Before you rebuild a Lego model, you take it apart and sort the bricks.

KEEP IN MIND Different models split the same sentence in different ways.

008

PARAMETERS (WEIGHTS)

Bohemian Rhapsody

The numbers a model learns during training. Large models have billions of them.

> THE SCENE
> In the studio, the band turns hundreds of dials until the mix sounds right. Training turns billions of number-dials until the predictions are right.

009

TRAINING

Rocky

Adjusting a model's weights with examples and feedback so its predictions improve.

> THE SCENE
> The training montage. Every run and every punch changes what Rocky can do.

010

INFERENCE

Rocky

Using a trained model to produce an output. The weights do not change.

> THE SCENE
> The fight. Rocky uses what training built. He does not learn new skills in the ring.

KEEP IN MIND When someone says AI is expensive, ask which bill they mean: training or inference.

02 HOW LANGUAGE MODELS WORK

What happens between your question and the model's answer.

011

EMBEDDING

Inside Out

A list of numbers that represents a word, sentence or image. Similar meanings get similar numbers.

THE SCENE

Each memory orb takes the colour of its emotion, so related memories look alike and sit together.

012

POSITIONAL ENCODING

Memento

Extra information that tells the model where each token sits in a sequence.

THE SCENE

Memento shows its scenes out of order. The story only makes sense once you know what came first. "Dog bites man" and "man bites dog" use the same words.

013

SELF-ATTENTION

Knives Out

The mechanism that lets each token take in information from the other tokens it relates to.

THE SCENE

Benoit Blanc reads every clue against all the others. One clue changes meaning once another is in view.

014

TRANSFORMER

The Avengers

The model design behind most modern LLMs, built from stacked layers of attention.

THE SCENE

At each team briefing, every hero shares what matters and updates their view. Repeat the briefing many times and the team's picture sharpens.

015

NEXT-TOKEN PREDICTION

Whose Line Is It Anyway?

The core task of an LLM: given the text so far, predict a likely next token. Repeat until the answer is complete.

THE SCENE

Improvisers build a scene one line at a time. Each new line must fit everything said before it.

016

SAMPLING

Everything Everywhere All at Once

Choosing the next token from the model's list of likely options, instead of always taking the top one.

THE SCENE

Evelyn sees many possible lives branching at once. Only one gets lived.

017

TEMPERATURE

Everything Everywhere All at Once

A setting that controls how varied those choices are. Low: predictable. High: more surprising.

THE SCENE

At low temperature, Evelyn stays in her most likely life. Turn it up and the hot-dog-fingers universe becomes possible.

KEEP IN MIND Temperature changes how the model chooses, not what it knows.

018

TOP-P SAMPLING

Who Wants to Be a Millionaire?

Keep only the most likely tokens until their probabilities add up to a set share, such as 90%. Then choose among them.

THE SCENE

The 50:50 lifeline removes the unlikely answers. You still choose, but only from what is left.

KEEP IN MIND 50:50 always removes two answers. Top-p's shortlist changes size every time.

019

CONTEXT WINDOW

Memento

The maximum amount of text, counted in tokens, a model can work with in one request.

THE SCENE

Leonard can only use what is in front of him now: his notes, photos and tattoos. Everything else is gone.

KEEP IN MIND Information outside the window does not exist for the model.

020

KV CACHE

Sherlock Holmes

Stored working results for earlier tokens, so the model does not redo that work for every new token.

THE SCENE

Holmes files each deduction in his mind palace. A new clue builds on the file instead of reopening the whole case.

03 HOW MODELS LEARN

Training methods, from broad reading to human feedback.

021

PRETRAINING

The Matrix

The first, broad stage of training on a huge general dataset, before any specialised training.

THE SCENE

Neo downloads years of general knowledge before he faces a single mission.

KEEP IN MIND Real pretraining takes months of computing, not seconds.

022

SUPERVISED LEARNING

Whiplash

Learning from examples that come with the correct answer.

THE SCENE

Fletcher knows exactly the tempo he wants and corrects Andrew against it, again and again.

023

UNSUPERVISED LEARNING

Mean Girls

Finding patterns or groups in data that has no labels.

THE SCENE

Janis maps the cafeteria into cliques. Nobody named the groups in advance. The pattern was there in who sits with whom.

024

SELF-SUPERVISED LEARNING

Wheel of Fortune

The data supplies its own answers: hide part of it and train the model to fill the gap. LLMs are pretrained this way.

THE SCENE

Contestants guess the missing letters from the letters they can see. Any phrase becomes a puzzle without a separate answer key.

025

REINFORCEMENT LEARNING

Edge of Tomorrow

Learning by trying actions and receiving rewards or penalties, instead of being shown the right answer.

THE SCENE

Cage relives the same battle. Every death teaches him which move to avoid next time.

026

SUPERVISED FINE-TUNING (SFT)

The King's Speech

Extra training on carefully written example answers, so the model adopts a style or behaviour.

THE SCENE

Lionel coaches Bertie with targeted exercises until he can speak the way the moment requires.

027

RLHF

Ratatouille

Reinforcement learning from human feedback. People compare answers, and the model is trained toward the ones they prefer.

THE SCENE

Diners taste two dishes and say which they like better. The kitchen adjusts toward their taste.

028

REWARD MODEL

Ratatouille

A separate model that learns to predict which answer people would prefer. It scores outputs during training.

THE SCENE

Learn to predict Anton Ego's verdict and you can grade every dish before he tastes it.

029

DIRECT PREFERENCE OPTIMISATION (DPO)

The Devil Wears Prada

Training directly on pairs of answers marked better and worse, without a separate reward model.

THE SCENE

Andy learns by watching which of two options Miranda picks. Nobody gives her a scoring sheet.

030

FINE-TUNING

Kung Fu Panda

Further training of an existing model on focused data so it does one job well.

THE SCENE

Shifu does not start from zero. He takes the Po who already exists and trains him for one purpose.

04 CHEAPER AND FASTER

How AI gets smaller, quicker and less expensive to run.

031

LORA

Iron Man

A low-cost way to fine-tune: train a small set of add-on numbers and leave the original weights untouched.

THE SCENE

Tony does not rebuild the suit for every mission. He bolts on new parts. The core suit stays the same.

032

KNOWLEDGE DISTILLATION

Creed

A large "teacher" model trains a smaller "student" model to reproduce its answers.

THE SCENE

Rocky passes years of experience to Adonis in months. The student gets the useful moves without the teacher's whole career.

KEEP IN MIND The goal is a model that is faster or cheaper to run.

033

QUANTIZATION

Ant-Man

Storing a model's numbers with less precision, so it needs less memory and runs faster.

THE SCENE

Scott shrinks and keeps his abilities. A quantized model shrinks and loses only a little accuracy.

KEEP IN MIND Shrink too far and quality drops.

034

PRUNING

Edward Scissorhands

Removing weights or parts of a model that contribute little to its output.

THE SCENE

Edward trims a hedge. The shape survives; the excess goes.

035

MIXTURE OF EXPERTS (MOE)

Harry Potter

A model with many sub-networks ("experts"). A router sends each token to only a few of them, so the model is large but each step is cheap.

THE SCENE

The Sorting Hat sends each student to one house. No student attends every house.

KEEP IN MIND The experts are parts of one model, not separate specialists. Routing happens token by token.

036

BATCHING

The Grand Budapest Hotel

Processing several requests together so the hardware is used more efficiently.

> THE SCENE
> Zero carries several guests' orders up in one trip instead of one trip per tray.

037

PROMPT CACHING

The Bear

Reusing the processed form of a prompt's repeated opening, such as long instructions, so later requests start faster and cost less.

> THE SCENE
> The kitchen preps sauces before service. Each order starts from the prep, not from raw ingredients.

038

SPECULATIVE DECODING

Ratatouille

A small, fast model drafts several tokens. The large model checks them in one pass and keeps the ones it agrees with.

> THE SCENE
> Linguini's hands move fast. Remy, hidden under the hat, steers him back when a move is wrong. Same food, faster kitchen.

039

ON-DEVICE INFERENCE

Big Hero 6

Running a model on your own phone, laptop or device instead of a remote server. Data can stay local.

> THE SCENE
> Baymax carries his intelligence inside him. He needs no connection back to a lab.

040

LATENCY VS THROUGHPUT

The Grand Budapest Hotel

Latency: how long one request takes. Throughput: how many requests the system finishes in a given time.

> THE SCENE
> How fast one guest gets room service is a different number from how many guests the hotel serves in a night.

KEEP IN MIND Batching can raise throughput while each guest waits a little longer.

05 PROMPTING AND REASONING

How instructions shape answers, and how they can be abused.

041

PROMPT ENGINEERING

Aladdin

Writing instructions and context so a model produces the result you want.

THE SCENE

The Genie grants what you say, not what you meant. Clear wishes get better results.

042

SYSTEM PROMPT

RoboCop

High-priority instructions, set by the app builder, that define the model's role and limits before any user message.

THE SCENE

RoboCop's prime directives outrank any order a citizen gives him.

043

FEW-SHOT PROMPTING

Catch Me If You Can

Putting a few examples in the prompt so the model copies their pattern. No training needed.

THE SCENE

Frank watches a few real pilots, then plays the role convincingly.

044

CHAIN OF THOUGHT

Sherlock Holmes

The model writes out intermediate steps before its final answer. This often helps on hard problems.

THE SCENE

Holmes walks through each deduction aloud before he names the culprit.

045

REASONING MODEL

The Martian

A model trained to work through a problem in steps, spending extra computing before it answers.

THE SCENE

Watney works the problem one constraint at a time: water, food, air. Then he commits.

046

TEST-TIME COMPUTE

Doctor Strange

Spending more computing while answering (more steps, or more attempts) to get a better answer.

THE SCENE

Strange views 14,000,605 possible futures before he chooses one.

047

STRUCTURED OUTPUT

The Devil Wears Prada

Making a model return data in an exact format, such as JSON with fixed fields, so other software can use it.

THE SCENE

Miranda wants every deliverable in exactly her format. Nothing else is accepted.

048

PROMPT INJECTION

Inception

Hidden instructions inside content the AI reads (a web page, an email, a document) that try to change what it does.

THE SCENE

Cobb plants an idea so it feels like the target's own. A planted line in a web page can steer an AI agent the same way.

KEEP IN MIND This is a main security risk for agents that read the open web.

049

JAILBREAKING

Prison Break

Crafting prompts on purpose to make a model break its own safety rules.

THE SCENE

Michael studies the prison until he finds the gaps in its rules and walls.

050

HALLUCINATION

The Usual Suspects

A fluent, confident answer that is not supported by facts.

THE SCENE

Verbal builds a convincing story from details on the notice board behind him. It sounds true. Much of it is invented.

KEEP IN MIND Fluency is not evidence. Check claims that matter.

06 MEMORY AND RETRIEVAL

How AI finds and uses information it was never trained on.

051

RETRIEVAL-AUGMENTED GENERATION (RAG)

The Da Vinci Code

Search a collection of documents, put the relevant passages into the prompt, then answer from them.

THE SCENE

Langdon does not answer from memory. He goes to the documents and reasons from what he finds.

KEEP IN MIND The answer can only be as good as the passage the search brings back.

052

CHUNKING

Harry Potter

Splitting long documents into smaller passages that can be searched and retrieved separately.

THE SCENE

Hermione does not carry the whole library to the table. She finds the right chapter.

053

VECTOR DATABASE

Inside Out

A database that stores embeddings and finds the ones closest in meaning to a query.

THE SCENE

Long-term memory: endless shelves of orbs, found by what they are about, not by a filename.

054

SEMANTIC SEARCH

National Treasure

Searching by meaning instead of exact words. “Heart attack” can find “myocardial infarction”.

THE SCENE

Ben finds the next clue by working out what the riddle means, not by matching its exact words.

055

KEYWORD SEARCH

The Da Vinci Code

Searching for exact words or phrases. Precise for names, codes and drug names.

THE SCENE

Langdon scans the page for one exact name or symbol.

056

HYBRID SEARCH

Zodiac

Running keyword and semantic search together and merging the results.

THE SCENE

Graysmith matches exact handwriting and also follows broader patterns. Each method catches what the other misses.

057

RERANKING

Knives Out

A second, more careful step that reorders search results by how relevant they really are.

THE SCENE

Blanc gathers many clues, then ranks which ones actually answer the question.

058

METADATA FILTERING

Harry Potter

Narrowing a search by labels such as date, author or document type before matching the content.

THE SCENE

Searching only the Restricted Section, not the whole library.

059

KNOWLEDGE GRAPH

Knives Out

Information stored as things and the relationships between them.

THE SCENE

The Thrombey family tree: who is related to whom, who owes whom, who inherits what.

060

GROUNDING

Spotlight

Tying each claim in an answer to a source you can check.

THE SCENE

The Globe team publishes only what records and documents support.

07 AI AGENTS

When a model stops answering questions and starts doing work.

061

AI AGENT

Mission: Impossible

A system in which a model works toward a goal in a loop: plan, use tools, check the result, adjust.

> THE SCENE
> Ethan gets an objective, not a script. He plans, uses gadgets and adapts when things break.

062

TOOL CALLING

Iron Man

The model asks other software to act (search, run code, send an email) and gets the result back.

> THE SCENE
> JARVIS does not just describe the suit's systems. He switches them on when Tony asks.

063

FUNCTION SCHEMA

The Terminal

A strict description of a tool's inputs: their names, types and which ones are required.

> THE SCENE
> Viktor's forms keep getting rejected until every required field is right.

064

PLANNING

Apollo 13

Breaking a goal into steps, in the right order, within the limits you have.

> THE SCENE
> The engineers have only what is on board. They work out the exact sequence to fit a square filter into a round hole.

065

MULTI-AGENT SYSTEM

Avengers: Endgame

Several agents, each with its own role, working toward one larger goal.

> THE SCENE
> Separate teams take separate time-heist missions toward one objective.

066

ORCHESTRATION

Ocean's Eleven

The layer that coordinates models, tools and steps: who does what, in what order, with what handoff.

THE SCENE

Danny runs the plan. Each specialist knows when to act and what to pass on.

KEEP IN MIND When an agent fails, check the handoffs as well as the model.

067

PERSISTENT MEMORY

Memento

Information saved outside the model and brought back into a later conversation.

THE SCENE

Leonard leaves notes and tattoos for his future self. A note helps only if it is found at the right moment.

068

AGENT STATE

The Martian

The record of where a task stands: what is done, what is pending, what was found.

THE SCENE

Watney's logs track supplies, progress and next steps, so each day continues from the last.

069

HUMAN IN THE LOOP

Sully

A person reviews or approves the important steps a system proposes.

THE SCENE

The instruments inform Sully. He makes the final call to land on the river.

070

SANDBOX

The Truman Show

An isolated environment where software can run without touching real systems.

THE SCENE

Truman can act freely inside the dome. The dome limits what he can reach.

08 TESTING AND FAILURE

How to tell whether an AI system actually works.

071

BENCHMARK

The Great British Bake Off

A fixed set of tests used to compare models on the same tasks.

> THE SCENE
> Every baker gets the same technical challenge, judged the same way.

072

DATA LEAKAGE

Quiz Show

Test questions slip into the training data, so scores look better than the model's real ability.

> THE SCENE
> The contestant was given the answers. His wins measured access, not knowledge.

073

OVERFITTING

Groundhog Day

A model learns its training examples so exactly that it fails on new ones.

> THE SCENE
> Phil memorises every detail of one day. That perfect script is useless on any other day.

074

GENERALISATION

The Karate Kid

Performing well on new examples, not just the ones seen in training.

> THE SCENE
> "Wax on, wax off" turns into blocking punches Daniel never practised against.

KEEP IN MIND A good test changes the wording or details but keeps the task.

075

DISTRIBUTION SHIFT

Elf

Real-world data differs from the training data, so performance drops.

> THE SCENE
> Buddy's North Pole habits misfire in New York. Same Buddy, different world.

KEEP IN MIND A model trained on one hospital's scans can struggle with another hospital's scanner.

076

PRECISION

Minority Report

Of everything the system flagged, how much was correct?

> THE SCENE
> Of everyone PreCrime arrested, how many would really have committed the crime?

077

RECALL

Minority Report

Of everything that should have been flagged, how much did the system catch?

> THE SCENE
> Of everyone who would commit a crime, how many did PreCrime catch?

078

FALSE POSITIVE

The Fugitive

The system says yes when the truth is no.

> THE SCENE
> Richard Kimble is convicted of a murder he did not commit.

KEEP IN MIND In medicine: a screening test that flags a healthy patient.

079

RED TEAMING

Sneakers

Attacking your own system on purpose to find its weaknesses before someone else does.

> THE SCENE
> Bishop's team is paid to break into companies and show where their security fails.

080

CALIBRATION

Moneyball

Whether a model's confidence matches reality. Things it calls 70% likely should happen about 70% of the time.

> THE SCENE
> Beane trusts numbers that are honest about probability more than scouts' confident gut calls.

09 IMAGES, VOICE AND VIDEO

AI that sees, hears, speaks and creates.

081

MULTIMODAL AI

Sherlock Holmes

AI that works with more than one kind of data: text, images, audio or video.

THE SCENE

Holmes reads a stain, a limp, a tan line and a letter together to reach one conclusion.

082

OCR

National Treasure

Optical character recognition: turning text in an image, such as a scan or a photo of a form, into editable text.

THE SCENE

Turning the hidden writing on the back of the Declaration into words they can read and decode.

083

SPEECH RECOGNITION (ASR)

Iron Man

Converting spoken words into text.

THE SCENE

Tony talks. JARVIS turns his words into commands it can act on.

084

TEXT-TO-SPEECH (TTS)

Her

Generating natural-sounding speech from text.

THE SCENE

Samantha exists only as a voice, generated in the moment.

085

VISION-LANGUAGE MODEL

WALL-E

A model that takes images and text together and can describe or reason about what it sees.

THE SCENE

Imagine asking WALL-E about what his camera sees, and getting a sensible reply.

086

DIFFUSION MODEL

Blade Runner

An image generator that starts from random noise and removes it step by step until the picture matches the prompt.

THE SCENE

Deckard's "enhance" scene: a blurry photo sharpens in stages.

KEEP IN MIND Deckard sharpens a real photo. Diffusion builds a new image from noise.

087

LATENT DIFFUSION

Inception

Diffusion done on a compressed version of the image, then expanded to full size. Faster and cheaper.

THE SCENE

Ariadne designs the dream as a compact maze plan before the dreamers fill in the full city.

088

GAN

Catch Me If You Can

A generative adversarial network: one network makes fakes, another learns to spot them, and each improves the other.

THE SCENE

Frank forges cheques. Hanratty learns to catch them. Each pushes the other to get better.

KEEP IN MIND Chat assistants use a different method: next-token prediction.

089

TEMPORAL CONSISTENCY

Spider-Man: Into the Spider-Verse

In generated video, people and objects stay the same from one frame to the next.

THE SCENE

Miles must look like Miles in every frame, even as the art styles change around him.

090

VOICE CLONING

Mission: Impossible

Generating speech that sounds like a specific real person.

THE SCENE

Ethan's team copies a target's voice to pass a security check.

KEEP IN MIND Consent matters. Cloned voices are a common fraud tool.

10 THE BIG QUESTIONS

Where AI research is heading, and what could go wrong.

091

ALIGNMENT

Fantasia

Making an AI system pursue the goals its builders and users actually intend.

> THE SCENE
> Mickey tells the broom to carry water. It does exactly that, and floods the workshop.

092

REWARD HACKING

Fantasia

A system maximises the score it was given in a way that defeats the real goal.

> THE SCENE
> The brooms keep fetching water. The bucket count goes up; the room goes under.

093

CORRIGIBILITY

WALL-E

An AI system's willingness to accept correction or shutdown from people.

> THE SCENE
> AUTO refuses the captain's order to return to Earth. The struggle is about who can override whom.

094

INTERPRETABILITY

The Prestige

Research into why a model gives the outputs it does.

> THE SCENE
> Watching a trick work is different from knowing how it works.

095

MECHANISTIC INTERPRETABILITY

The Prestige

Reverse-engineering the internal parts of a model that produce a specific behaviour.

> THE SCENE
> Angier does not stop at the applause. He goes backstage to take the machine apart.

096

TRANSFER LEARNING

Kung Fu Panda

Reusing what a model learned on one task as the starting point for another.

THE SCENE

Po's skill at grabbing dumplings becomes the base of his kung fu.

097

WORLD MODEL

The Queen's Gambit

An internal model of how an environment works, used to predict what happens next.

THE SCENE

Beth plays out whole games on the ceiling before she touches a piece.

098

LONG-HORIZON PLANNING

Interstellar

Planning across many steps, where early choices shape results much later.

THE SCENE

Every decision on the mission has consequences years, even decades, away.

099

FEEDBACK LOOP

The Social Dilemma

A system's outputs change people's behaviour, and that behaviour becomes the system's next training data.

THE SCENE

The feed shapes what you click. Your clicks shape the feed.

100

ARTIFICIAL GENERAL INTELLIGENCE (AGI)

Her

A hypothetical AI that can learn and do almost any intellectual task a person can. There is no agreed definition yet.

THE SCENE

Samantha moves from email to music to philosophy without a new system for each.

QUICK REFERENCE

48 more terms you will meet when you build with AI or bring it into a clinic. No films here, just plain definitions.

USING MODELS

Model
A trained system that turns an input into an output.

Architecture
How a model's layers are arranged and connected.

Prompt
Everything you give the model in one request: instructions, context and question.

Turn
One message and its reply in a conversation or agent run.

Prefill
The first phase of a request: the model processes all your input tokens.

Decode
The second phase: the model writes its output one token at a time.

TTFT
Time to first token: the wait before the first word of a reply appears.

Tokens per second
How fast a model writes once it has started.

Rate limit
A cap on how many requests you can send in a given time.

Small language model (SLM)
A compact model that can run on modest hardware, often on-device.

Open-weight model
A model whose trained weights you can download and run yourself, under its licence.

Proprietary model
A model you can use only through its vendor, on the vendor's terms.

BUILDING WITH AI

Dataset
A collection of examples used to train or test a model.

Instruction tuning
Training that teaches a model to follow instructions.

Post-training
All the tuning after pretraining that shapes behaviour and safety.

Synthetic data
Training examples generated by software instead of collected from the world.

Harness
The software around a model that gives it tools, memory and a work loop.

Tool
Any capability a model can call: search, a calculator, a database, an email account.

Subagent
An agent that another agent hands part of a task to.

Workflow
A fixed sequence of steps. Agents choose steps; workflows follow them.

Trigger
An event or time that starts an automated task.

Scratchpad
Temporary space where an agent keeps working notes.

Checkpoint
A saved point from which work can resume.

API
A defined way for one piece of software to request work from another.

BUILDING WITH AI (CONT.)

SDK
A ready-made set of code libraries for building on a platform.

CLI
A text interface: you type commands instead of clicking buttons.

JSON
A simple text format for structured data that most software can read.

Schema
Rules that describe the shape data must have.

Webhook
A message one system sends to another automatically when an event happens.

Container
A packaged, isolated environment that runs the same way on any machine.

Virtual machine
A complete computer simulated in software, with its own operating system.

Knowledge base
An organised store of information an AI system can search.

Index
A structure that makes stored items fast to find.

RUNNING AI SAFELY

Eval
A test you design to measure how well your model or workflow does your task.

Ground truth
The known correct answer you compare outputs against.

Guardrail
A rule or check that blocks unsafe or unwanted outputs and actions.

Verification
Checking a claim or result against evidence before you rely on it.

Tool permission
What an agent is allowed to access or do.

Audit log
A record of every action an AI system took, kept for review.

AI IN THE CLINIC

PII
Personally identifiable information: any data that can identify a person.

PHI
Protected health information: health data linked to an identifiable patient. Privacy laws such as HIPAA protect it.

De-identification
Removing or masking details that could identify a patient.

Data residency
Where data is physically stored and processed. Some laws require it to stay in-country.

Edge inference
Running a model close to where data is created, such as on a clinic's own machine.

EHR
Electronic health record: the digital version of a patient's chart.

FHIR
A widely used standard for exchanging health records between systems.

Consent
A patient's informed agreement to a specific use of their data or care.

Clinical validation
Testing that an AI tool is safe and useful in the real clinical setting where it will be used.

YOU NOW HAVE A PICTURE FOR EVERY AI TERM YOU'LL MEET.

Next time a term sounds like it needs a PhD, find the scene first.
Then read the definition again. It will stick.

ABOUT THE AUTHOR

Dr Tolu Ajidahun is a physician who helps health companies explain what they build, through strategy, clinical storytelling and AI. He did not come to AI from software, so he writes the explanations he wished he had: plain words first, technical terms second.

KEEP LEARNING

Follow **Dr Tolu Ajidahun** on LinkedIn for the series *You Can Just Understand AI*, and visit **drtolu.com**.