Every AI Term Explained

The 52 terms from the film, in the same six groups and the same order. Every definition here is the one used on screen. Where the common explanation is wrong, the correction is in bold.

52 TERMS · 6 GROUPS · IDEASREPAY.COM/ACADEMY/AI-TERMS

1 Foundations

13 terms

01
Agentic AI
Software given a goal instead of a question, and allowed to take actions until it reaches it. A real shift, and also the pitch that raises money.

02
Generative AI
Models that produce new content: text, images, audio, video, code. Not a rival to agentic. Generative is the engine; agentic is what you let it drive.

03
AI: the parent branch
The whole field, dating to the 1950s. Machine learning sits inside it, deep learning inside that, and generative AI inside that.

04
Rules / if-else
Software where a human writes every condition in advance. You always know why it did what it did, and it cannot handle a case nobody thought of.

05
Machine learning
Instead of writing the rules, you show thousands of examples and the system works the rule out. Still runs fraud detection and spam filtering today.

06
Train-test split
Hold part of the data back and test only on that. Doing well only on what it trained on is overfitting. **80/20 is a habit, not a standard: scikit-learn defaults to 75/25.**

07
Classification & regression
Classification predicts a category (spam or not spam). Regression predicts a number on a continuous scale. A discrete class, or a continuous value.

08
Neural networks
Layers of simple units joined by weighted connections. Training nudges the weights until the answers stop being wrong. Arithmetic at enormous scale.

09
Deep learning
A neural network with more than one hidden layer. That is the actual definition: more than one, not dozens. It took off when GPUs and datasets caught up with an old idea.

10
Computer vision
Images and video. Detection draws a box, segmentation labels every pixel, pose estimation finds keypoints. The field's oldest commercial success.

11
Transformers
The June 2017 architecture from "Attention Is All You Need". Takes the whole sequence at once and weighs every word against every other. The T in GPT.

12
Applied AI
The shift from "can this work in a lab" to "is anyone shipping it". Most of the money now is in application, not invention.

13
Open weights vs open source
Open weights: download it, run it, change it. Open source (OSI definition, 2024) also wants the code and enough data information to rebuild an equivalent. **It does not demand the dataset itself.** Gemma is open weights.

2 The Model Itself

9 terms

14
LLM
Large language model. It reads what came before and predicts what comes next. Everything else is scaffolding built around that one move.

15
Tokens
Chunks of text, not words. You are billed per token and every limit is counted in them. **The "4 characters per token" rule is OpenAI-specific, not universal, and it drifts as tokenizers change.**

16
Next-token prediction
Produce a probability for every token that could come next, pick one, append it, and run again. Thousands of times.

17
Temperature
The dial on that pick. Low is predictable, high flattens the odds. **Ranges differ: Anthropic 0 to 1, OpenAI and Google 0 to 2. And temperature 0 is not deterministic**, because other users' traffic changes your batch size and so the order of the floating-point sums.

18
Context window
Everything the model can see at once. Not memory: a desk with an edge. **1M tokens is now the default on top models, but accuracy and recall still degrade as it fills. The name for that is context rot, coined by Chroma; Anthropic's docs adopted it, crediting Chroma, in September 2025.**

19
Hallucination
Defined relative to the source, not the world: output that is nonsensical or unfaithful to what it was given. Intrinsic contradicts the source; extrinsic cannot be checked against it, and can happen to be true.

20
Probabilistic vs deterministic
Deterministic: same input, same output, every time. Probabilistic: same input, likely output. Use code for facts and the model for judgement.

21
Local / small models
Cheap, fast, often good enough, and some run entirely on your own hardware at zero cost per call. The skill is knowing the smallest model that still passes.

22
Tool calling reliability
Whether the model picks the right function and fills the arguments correctly. **The top score on Berkeley's function-calling leaderboard is about 77%.** Roughly one call in four is wrong at the top of the board.

3 Agents
8 terms

23
Chatbot vs AI agent
A chatbot answers and waits. An agent pursues a goal across steps without asking between each one. Its reach is only what you granted it.

24
Core agent loop
Observe, reason, act, feedback. Google's glossary names those four stages and says the cycle repeats "until a termination condition is met". Note it is reason, not think. That loop is the entire difference between a chatbot and an agent.

25
ReAct pattern
Reasoning and acting, interleaved (October 2022). The model states what it is about to do and why, does it, then reads the result. The trace is loggable.

26
Tools
A function plus a description of what it does and what arguments it takes. The model never runs your code: it asks by name and your system executes.

27
Observability
A replayable record of every prompt, tool call, argument, response, token and latency. Agents fail silently far more often than they crash.

28
Sandboxing
Running untrusted code in an environment whose permissions are cut to the essential set. Not because the model is malicious, but because it is confident.

29
Human in the loop
The agent must stop and ask before certain actions. **The EU AI Act names a stop button and automation bias, for high-risk systems specifically.**

30
Forward-deployed engineer
An engineer embedded with the customer, building in their environment with their data. From Palantir, now a standard AI job title.

4 Knowledge and Data
10 terms

31
Memory (amnesia by default)
The model remembers nothing. A chat only seems to because the conversation is resent every single turn. Not always whole: compaction and context editing trim what goes back. Stateless either way, so memory is a design problem.

32
Working memory
The current task and session. It lives in the context window and dies with it.

33
Episodic memory
Specific events with a time attached, stored outside the model and pulled back when relevant. The kind that makes an assistant feel like it knows you.

34
External database memory
Facts written to an ordinary database and read back on the next call. If you know exactly what to look up, a database row beats anything clever.

35
Graph memory
Stores relationships rather than documents, so an answer can be traced along a path across several facts. The good ones also store when a fact was true.

36
RAG
Retrieval augmented generation (Facebook AI Research, UCL and NYU, 2020). Search your own documents first and put the passages that matter into the prompt.

37
Chunking
How you cut documents before retrieval. **Strategies differ by up to 9% in recall, and the widely-copied default scores below average.** Most bad RAG is bad chunking.

38
Vectors & dot product
Text as a long list of numbers placing it in space. Magnitude is length, direction carries meaning. Compare with the dot product. **Cross product does not apply: it exists only in three dimensions and returns a vector, not a rankable score.**

39
Meaning vs fuzzy search
Fuzzy search forgives typos but still matches letters. Meaning-based search matches sense, with no shared words needed.

40
Vector databases
Stores built to hold millions of vectors and return the nearest in milliseconds, without comparing against every one. Qdrant, Pinecone, Postgres with an extension.

5 Protocols and Architecture
9 terms

41
MCP

Model Context Protocol. One standard way for a model to reach a tool or data source. Announced by Anthropic in November 2024; **a Linux Foundation project since December 2025.**

42
A2A

Agent-to-agent protocol. Google, April 2025; Linux Foundation, June 2025. **By its own documentation, not a replacement for MCP but complementary to it.**

43
Smart agent architecture

A grand phrase for a plain truth: there is no standard shape. Learn the pieces, then assemble for your own problem.

44
Chain of thought

Working through the steps before answering. Eight worked examples in the prompt beat a model fine-tuned for the job. **The written reasoning is more predicted text, not a transcript of the machine.**

45
Plan and execute

Produce the plan, then work the plan, with the plan written down as a list. One place to look when it fails, one place to step in.

46
Evals

Tests for something that does not give the same answer twice. Graded by rule or by another model. Prefer volume over polish.

47
Multi-agent systems

One job split across several agents. Anthropic measured a 90% gain on their research eval, at roughly 15x the tokens of a chat, and it works far less well for coding. **"MASA" is not a standard term.**

48
Manager & worker agents

A manager holds the goal and hands out pieces; workers do one narrow thing and report back. Failures cluster on the report, not on the work.

49
Git worktrees

Multiple working trees attached to one repository, so you can check out more than one branch at a time. By convention, how several coding agents work without colliding.

6 Safety and Economics

3 terms

50
Guardrails & safety

Checks on the way in (prompt injection is the real threat) and on the way out (structure, leaks, tone). NVIDIA's NeMo Guardrails is an open toolkit; AWS Bedrock Guardrails is a managed service, claiming to block **up to** 88% of harmful content.

51
Prompt vs middle-layer guardrails

A rule written into the prompt travels the same channel as the attack, so it can be overridden. A code layer outside the model cannot be talked out of it. Prompt for tone, layer for harm.

52
Cost management

Tier the work: cheap models for simple calls, the expensive one only for the hard tenth. **The 60-30-10 rule is folk guidance, published by nobody in AI.** The citable version is routing research, which cuts cost by more than half without losing quality.

The thing worth keeping

- Almost every term on this sheet is the name of a **workaround**. Memory exists because the model forgets. RAG exists because it does not know your data. Guardrails exist because it will say anything. Evals exist because it will not repeat itself. Tools exist because on its own it cannot act at all.
- None of it is magic. It is scaffolding built around a next-token predictor.
- And one number for scale: the widely quoted agent time-horizon figures are the **50%** ones, which is a coin flip. At **80%** reliability, the level you would need to actually ship something, the best models measured top out at around an hour to ninety minutes of expert human work.