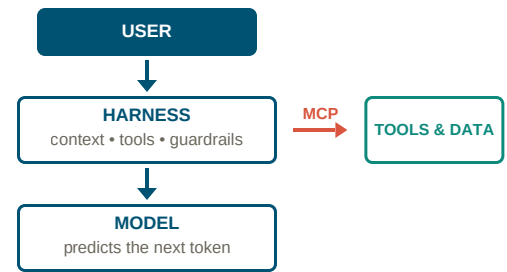


The AI Foundations Cheat Sheet

Essential AI Terms that Technical Professionals Need to Know

THE MODERN AI STACK



INSIDE THE MODEL

LLM - Large Language Model

A neural network trained on huge amounts of text to predict the next token. Knowledge is often very broad.

Open Weight vs. Closed Weight

Open weight is an AI model whose parameters are public for anyone to download and run. Closed weight stays private and is accessed through an API or interface.

Parameters

The learned weights inside a model. More parameters generally means more capability - and more cost. Often seen as part of the file (70B, etc.)

Training vs. Inference

Training builds the model's weights up front by a lab/organization - inference is the model answering you at run time.

Token

The small chunk of text a model reads and writes - roughly 4 characters. Usage is billed per token, in and out. Use a tokenizer to see visually.

Context Window

The model's working memory - the maximum tokens it can process at once. What doesn't fit here will get left out of consideration.

Temperature

A sampling setting. Low temp keeps output focused and repeatable, high temp makes it varied and creative.

Hallucination

A confident but false answer. Models predict likely-sounding text - they don't verify facts before outputting them.

WORKING WITH MODELS

Prompt Engineering

Structuring the instructions - such as adding a role, examples, constraints, output format - so the model returns reliable results.

System Prompt

Instructions the model receives before your message, setting its behavior and rules. Often part of the harness and may not be seen by users.

Context Engineering

Managing everything in the window - instructions, docs, history, and tools - not just the prompt.

RAG - Retrieval-Augmented Generation

Fetch relevant documents at question time and add them to the prompt so answers stay grounded in your data.

Embedding

Text converted to a vector of numbers where similar meanings land close together. Powers semantic search.

Fine-Tuning

Extra training that adds new behavior into the model's weights. RAG adds knowledge; fine-tuning changes the model.

AGENTS & THE ECOSYSTEM

Agent

AI that works a goal in a loop - reason, act, observe - instead of answering in a single shot.

Tool Calling

The model outputs a structured request and the software around it runs the tool and returns the result. Examples include search, Bash, Python, etc.

Harness

Everything around the model - context assembly, tools, guardrails, approvals. The other half of the product.

Subagent

A helper agent spawned for one subtask, with its own context and instructions, reporting back when done.

Orchestrator

Coordinates multi-step, multi-agent workflows - the role frameworks like LangChain and LangGraph play.

Human in the Loop

Checkpoints where a person approves an agent's plan or action before it runs - especially the risky ones.

MCP - Model Context Protocol

An open standard that connects agents to tools and data - one protocol instead of custom integrations everywhere.

Agent Skills

Packaged instructions that teach an agent your workflow. MCP gives agents hands; skills give them your playbook.

AI RUNTIME SECURITY

Prompt Injection

Malicious instructions hidden in content the agent reads - a ticket, doc, or web page - that hijack its behavior.

Least Privilege

Grant an agent only the access each task needs. Over-privileged agents are the top governance risk.

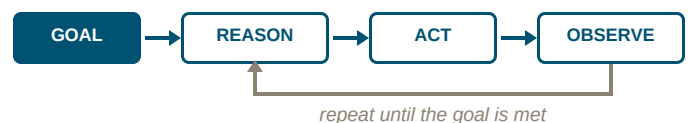
Observability

Logging what the AI saw, decided, and did - so you can audit, debug, and trust it.

Identity

A unique ID or profile that allows the agent to prove who it is when accessing APIs, data, or tools. Also allows identification in logs and auditing.

THE AGENT LOOP



Human-in-the-loop: tasks that required approval for risky actions