Glossary

Term	Simple Explanation
Agent	An AI system that can reason, use tools, and perform tasks toward a goal.
Agent Harness	Software that repeatedly calls an LLM, performs actions, and validates results until a task is complete.
Cached Tokens	Previously processed context that can be reused by the model.
Context	All information available to the AI when generating a response.
Context Engineering	Providing the right information, tools, and rules to help AI perform effectively.
Context Window	The maximum amount of information an AI can consider at one time.
Custom Agent	A specialized AI agent with specific instructions, tools, or expertise.
Embedding	A numerical representation of meaning that helps AI compare related content.
Fine-Tuning	Additional training of a model for specific tasks or domains.
Hallucination	When AI generates information that sounds correct but is actually wrong.
Handoff	A structured summary that transfers context between AI sessions.
Harness Engineering	Designing the harness around the model – tools, context, verification and validation – so an agent performs reliably on real tasks.
Hook	An automated, deterministic script triggered at specific events.
Inference	The process of generating an answer using a trained model.
Input Tokens	Tokens sent to the model as input.
LLM (Large Language Model)	An AI model trained to predict and generate text.
Loop Engineering	Loop engineering is a shift from manual, single-shot AI prompting to building autonomous, self-correcting, continuous execution cycles.
MCP (Model Context Protocol)	A standard for making external systems and APIs available as AI tools.
Non-Deterministic	Producing different outputs for the same input on different runs.
Output Tokens	Tokens generated by the model as a response.
Progressive Disclosure	Revealing context, instructions or tools to an agent only when they are needed, instead of loading everything up front.
Prompt	The instruction or question given to an AI.
Prompt Engineering	Crafting prompts to get better results from AI.
RAG (Retrieval-Augmented Generation)	A technique where AI retrieves external information before answering.
Research → Plan → Implement	A structured workflow that reduces mistakes and improves results.
Skill	A reusable package of knowledge, workflows, or capabilities.
Spec-Driven Development (SDD)	Building software from specifications before writing code.
Subagent	A secondary agent handling a specific subtask independently.
System Prompt	Instructions that define how the AI should behave.
Temperature	A setting that controls how creative or predictable an AI response is.
Token	A small unit of text, such as a word or part of a word.
Tool	A capability that lets AI interact with external systems or perform actions.
User Prompt	The request entered by the user.
Validation	Checking whether the solution solves the real problem. (“Are we building the right thing?”)
Vector	A list of numbers representing the meaning of data.
Verification	Checking whether the solution was built correctly. (“Are we building it right?”)