

# The AI Stack

A layered architecture guide to the modern AI ecosystem

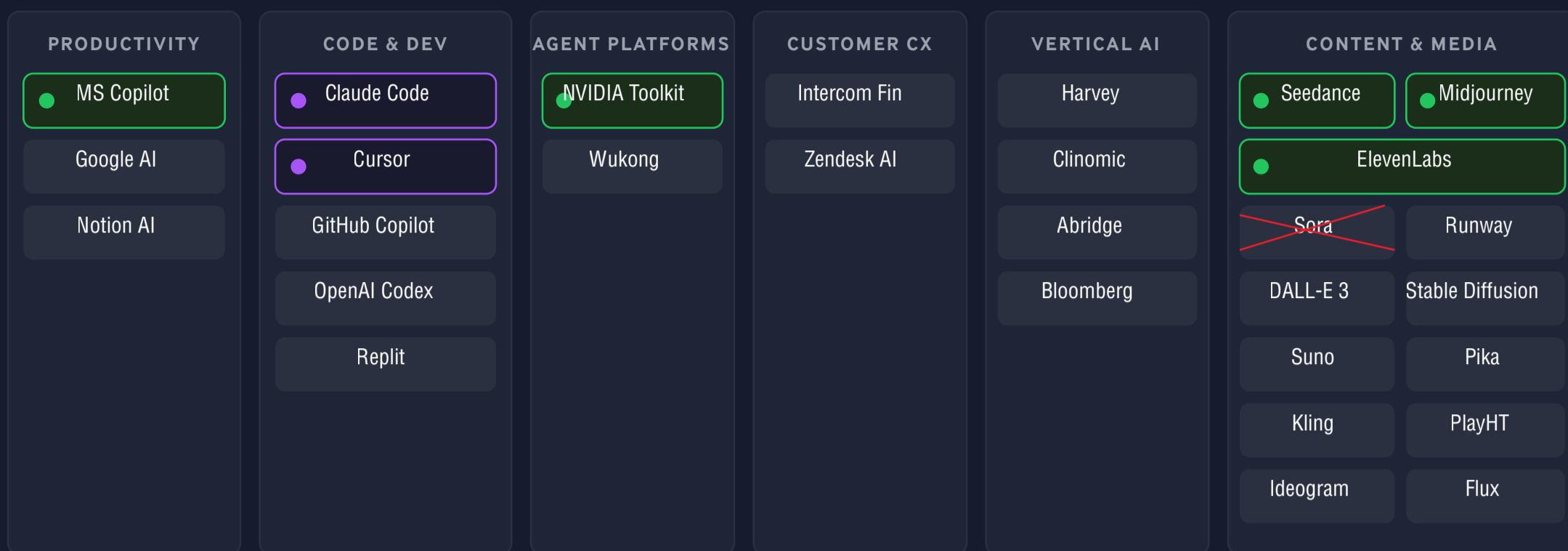
Updated March 2026 | 90+ Companies | 8 Layers

Read from the bottom up. Each layer builds on the one below.

## LAYER 08

### Applications & Agents

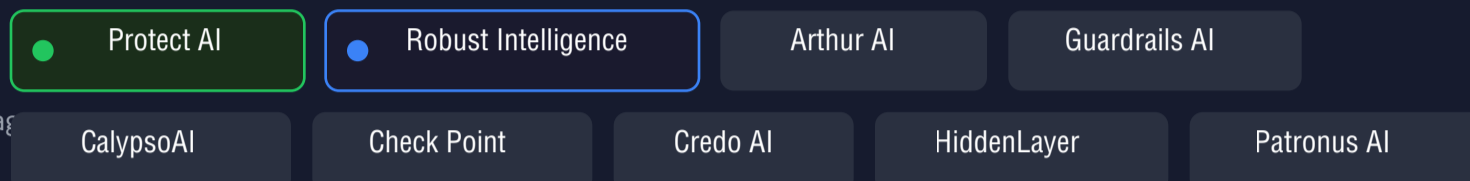
End-user products, autonomous systems & content generation



## LAYER 07

### Safety & Governance

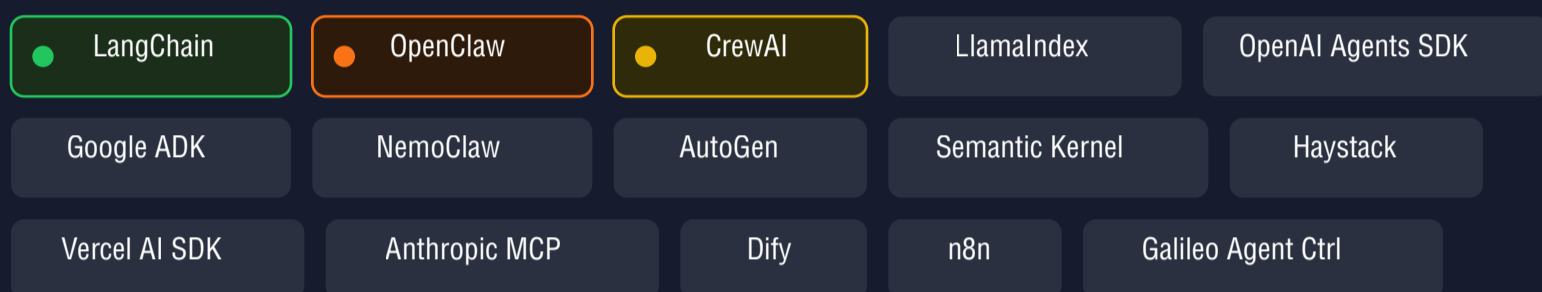
Red-teaming, guardrails, compliance & risk management



## LAYER 06

### Orchestration & Agents

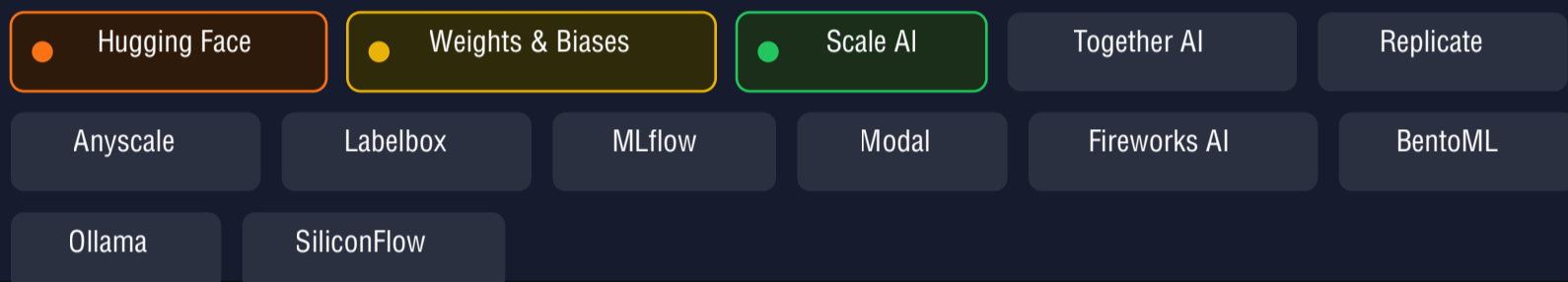
Frameworks, agent toolkits & workflow engines



## LAYER 05

### Model Dev & MLOps

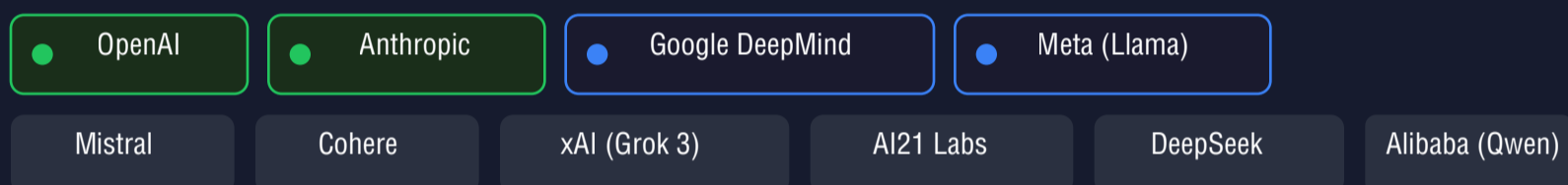
Training, fine-tuning, evaluation & deployment



## LAYER 04

### Foundation Models

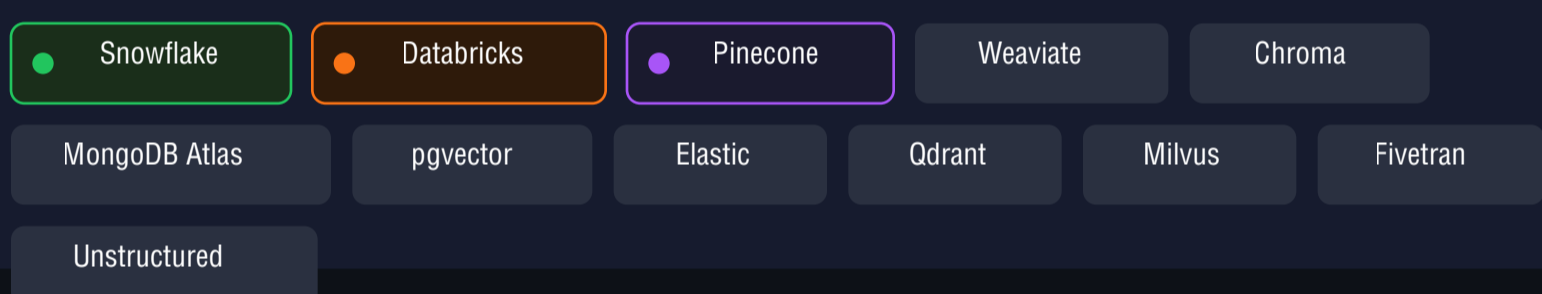
Large language, vision & multimodal models



## LAYER 03

### Data & Vector Storage

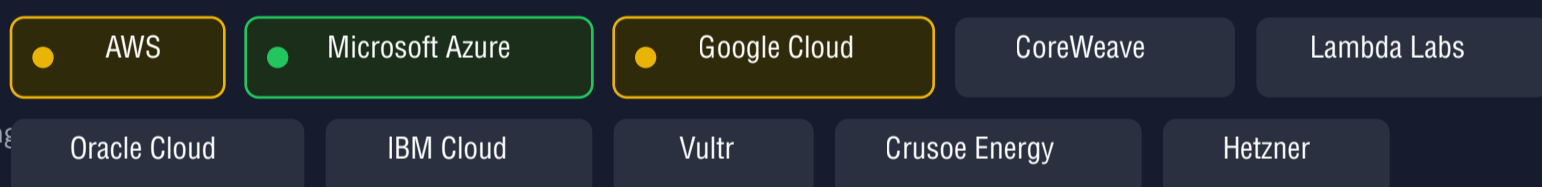
Databases, vector stores, data pipelines & ETL



## LAYER 02

### Cloud & Compute

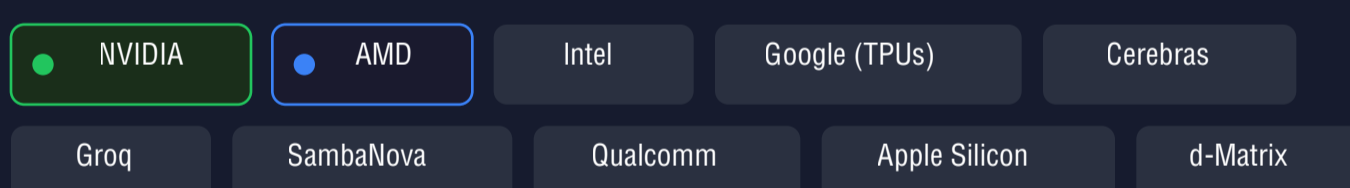
Cloud platforms, GPU clusters & inference hosting



## LAYER 01

### Silicon & Hardware

GPUs, TPUs, custom silicon & accelerators



Green - Primary Leader / Purple = Breakout Leader / Blue = Secondary Major Leader / Orange = Important Challenger / Yellow = Significant Player

## How Executives Should Read This

### BUILD VS. BUY

Most companies should buy at Layers 1-4 and build differentiation at Layers 5-8.

### SECURITY FIRST

Layer 7 is not optional. Every production AI deployment needs guardrails and red-teaming.

### AGENT ERA

2026 is the year of autonomous agents. Layer 6 frameworks enable AI that acts, not just answers.

### AVOID LOCK-IN

Use open standards (MCP, ONNX) at orchestration layers. Swap models without rewriting apps.

### VALUE ACCRUAL

Value is shifting up the stack. Application layer companies capture the most end-user revenue.

### TCO REALITY

GPU costs dominate. Plan for inference spend to exceed training spend as adoption scales.

## 15 Terms Every Executive Needs

### LLM

Large Language Model. The neural network that powers text generation.

### Fine-Tuning

Customizing a pre-trained model on your domain data.

### Tokens

Units of text (roughly words/subwords). Usage is priced per token.

### Prompt Engineering

Crafting inputs to get better outputs from AI models.

### Inference

Running a trained model to produce predictions or text.

### Multimodal

Models that process text, images, audio, and video together.

### Foundation Model

A large pre-trained model that serves as a base for applications.

### Open vs. Closed

Open-weight models (Llama) vs proprietary APIs (GPT, Claude).

### RAG

Retrieval-Augmented Generation. Feeding relevant docs to the model.

### Vector Database

Stores embeddings for semantic search and retrieval.

### AI Agent

An AI system that can plan, use tools, and act autonomously.

### Hallucination

When the model confidently generates false information.

### MCP

Model Context Protocol. Standard for connecting AI to tools/data.

### Guardrails

Safety filters that prevent harmful or off-topic model outputs.

### GPU

Graphics Processing Unit. The hardware that powers AI training.

LAYER 04

# Foundation Models

Large language, vision & multimodal models

● OpenAI

● Anthropic

● Google DeepMind

● Meta (Llama)

Mistral

Cohere

xAI (Grok 3)

AI21 Labs

DeepSeek

Alibaba (Qwen)