

The AI Stack

A layered architecture guide to the modern AI ecosystem

Updated July 2026 | 170+ Companies | 8 Standards | 9 Layers | 4 Buying Motions

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

BUSINESS APPLICATIONS Where value accrues and where buying should start.

LAYER 09

Applications & Agents

End-user products, autonomous systems, content generation and the consoles that govern agent fleets

AGENT PLANES **NEW**

OpenAI Frontier · S

Microsoft Agent 365 · S*

Agentforce 360 · S*

Gemini Enterprise · S

Glean Agents · S

PRODUCTIVITY

Microsoft 365 Copilot · S

Glean · S

Gemini for Google Workspace · S

Notion AI · S

Claude Cowork · S

Writer · S

Moveworks · S

CODE AND DEV

Claude Code · S

Cursor · S*

GitHub Copilot · S

OpenAI Codex · S

Cognition Devin · S

Replit · S

Lovable · S

Base44 · S

CUSTOMER CX

Sierra · S

Decagon · S

Intercom Fin · S

Zendesk AI · S

Cognigy · S

Kore.ai · S

Parloa · S

VERTICAL AI

Harvey · S

Abridge · S

OpenEvidence · S

Legora · S

Ambience Healthcare · S

Hippocratic AI · S

Rogo · S

Hebbia · S

CONTENT AND MEDIA

Kling · S

Seedance · S*

ElevenLabs · S

Veo · S

Runway · S

Midjourney · S

FLUX · H

Ideogram · S

Suno · S

MiniMax Hailuo · S

Alibaba Wan · H

SYSTEMS OF RECORD

Already in your contract. The question is how much of their AI you turn on.

Salesforce · S

ServiceNow · S

SAP · S

Workday · S

HubSpot · S

Atlassian · S

Intuit · S

RUNTIME CONTROL AND GOVERNANCE What agents are allowed to do, and how you prove it afterward. Now drives cost and risk as much as the model layer.

LAYER 08

Security, Safety & Governance

Guardrails, agent identity, red-teaming, compliance and regulatory risk

CATEGORY LEADERS

Palo Alto Prisma AIRS · S/E

Check Point AI Agent Security · S

Microsoft Entra Agent ID · S

NOTABLE PLAYERS

Noma Security · S

HiddenLayer · S/E

CrowdStrike · S

Cisco AI Defense · S

SentinelOne · S

F5 AI Guardrails · E*

Oasis Security · S

Astrix Security · S

XBOW · S

WitnessAI · S

Mindgard · S

STANDARDS AND GUIDANCE

[OWASP TOP 10 FOR AGENTIC APPLICATIONS 2026] · F

[NIST AI RISK MANAGEMENT FRAMEWORK] · F

[ISO/IEC 42001] · F

[EU AI ACT / EU AI OFFICE] · F

[ANTHROPIC RESPONSIBLE SCALING POLICY] · F

LAYER 07

Orchestration & Protocols

Frameworks, agent toolkits, workflow engines and the interop protocols underneath them

CATEGORY LEADERS

LangGraph · H

OpenAI Agents SDK · H

Claude Agent SDK · H

NOTABLE PLAYERS

Agent Development Kit (ADK) · H

Microsoft Agent Framework · H

Temporal · H/S ■

n8n · H/S ■

Dify · H

CrewAI · H

LlamaIndex · H

Pydantic AI · H

NVIDIA NeMo Agent Toolkit · H

OpenClaw · H

Amazon Bedrock AgentCore · S

Foundry Agent Service · S

Vapi (voice) · S

STANDARDS AND GUIDANCE

[MODEL CONTEXT PROTOCOL (MCP)] · P

[AGENT-TO-AGENT (A2A)] · P

OPENCLAW CAUTION:

Roughly 368,000 stars and 12M downloads. Also five formal CVEs including CVE-2026-32922 (CVSS 9.9, paired-device escalation to admin RCE), the ClawHavoc supply-chain campaign, and 135,000+ exposed instances across 82 countries with roughly 63 percent running no authentication. Deploy hardened or not at all.

DATA AND MODEL STACK Where switching cost accumulates. Data gravity is permanent, model choice should not be.

LAYER 06

Model Dev, Evals & Observability

Fine-tuning, evaluation, tracing and production monitoring for models and agents

CATEGORY LEADERS

Hugging Face · S/H ■

LangSmith · S/E

Braintrust · S

NOTABLE PLAYERS

Langfuse · H

Phoenix · H

Arize AX · S

W&B Weave · S

MLflow · H

Databricks Mosaic AI · S

Galileo · S

Promptfoo · H

Patronus AI · S

DeepEval · H

FINE-TUNING

Training and adaptation, not tracing or evaluation.

Axolotl · H

Unsloth · H

STANDARDS AND GUIDANCE

[OPENTELEMETRY GENAI SEMANTIC CONVENTIONS] · F

LAYER 05

Foundation Models

Large language, vision and multimodal models, closed and open weight

CATEGORY LEADERS

Anthropic · S

OpenAI · S

Google DeepMind · S

NOTABLE PLAYERS

xAI · S

Meta (Llama) · H*

Mistral · H/S ■

DeepSeek · H

Alibaba (Qwen) · H/S* ■

Moonshot (Kimi) · H

Z.ai (GLM) · H

MiniMax · H/S ■

Cohere · S/E

Sakana · S

Poolside (code) · S/H

LAYER 04

Data & Context Infrastructure

Databases, vector stores, pipelines and the governed context agents relieve at runtime

CATEGORY LEADERS

Databricks · S/E

Snowflake · S

PostgreSQL + pgvector · H

NOTABLE PLAYERS

Pinecone · S

Qdrant · H/S ■

Weaviate · H/S ■

Milvus · H*

Chroma · H

LanceDB · H

Turbopuffer · S

Neon · S*

Crunchy Data · E*

Fivetran · S

dbt · H/S* ■

Atlan · S

INFRASTRUCTURE SPEND Power, capacity and tokens. Metered, volatile, and the largest line item.

LAYER 03 **NEW**

Inference & Model Serving

Serving models at production scale and cost. The layer that sets your gross margin

CATEGORY LEADERS

Fireworks AI · S/E

Together AI · S

Baseten · S/E

NOTABLE PLAYERS

Modal · S

Groq · S†

Cerebras · S†

Replicate · S

vLLM · H

SGLang · H

Ollama · H

OpenRouter · S

LiteLLM · H ■

Portkey · S/H

Amazon Bedrock · S

Microsoft Foundry Models · S*

Google Vertex AI · S

LAYER 02

Cloud, Capacity & Power

Cloud platforms, GPU clusters, neoclouds and the electricity that now constrains all of it

CATEGORY LEADERS

AWS · C

Microsoft Azure · C

Google Cloud · C

NOTABLE PLAYERS

CoreWeave · C

Nebius · C

Oracle Cloud Infrastructure · C

Lambda · C

Crusoe · C

Nscale · C

Fluidstack · C

Meta Compute · C*

ADJACENT SUPPLY

Regional hosting and power supply. Different buying motion from the neoclouds above.

Hetzner · C

IREN · C

LAYER 01

Silicon & Hardware

GPUs, TPUs, custom silicon and inference accelerators

CATEGORY LEADERS

NVIDIA · C

AMD · C

Broadcom · C

NOTABLE PLAYERS

Google (TPU) · C*

AWS Trainium · C

Marvell · C

Groq · C‡

Cerebras · C‡

Microsoft Maia · C

Meta MTIA · C

MediaTek · C

Qualcomm · C

Apple Silicon · C*

Hailo · C

Intel · C*

WHAT IT IS

Vendor

Commercial vendor

oss

Open source project or protocol

[TEXT]

Standard, framework or guidance body

oss + ■

OSS core, commercial offering available

HOW YOU GET IT (pffer the · on each chip)

S

SaaS or managed API

H

Self-hostable

E

On-prem or VPC capable

P

Protocol

F

Standard or framework

C

Silicon or physical capacity

HOW TO READ IT

Solid fill

Category leader

Outline

Notable player

Buying motion, see left rail

Name not verified against current public product name

Acquired, parent listed in this layer

Operates in two layers, listed in both

Read bottom to top. Entity type tells you what you are buying. Color tells you which budget it comes out of.

How Executives Should Read This

BUILD VS. BUY

Buy Layers 1 through 5. Build your differentiation at Layers 7 through 9. Nobody wins by owning infrastructure.

SECURITY FIRST

Layer 8 is not optional. The unsolved problem is agent identity; systems built for humans handing credentials to non-humans.

AGENT ERA

Nearly every enterprise now runs agents in production. Far fewer can govern them. The constraint stopped being capability and became control.

POWER IS THE CONSTRAINT

Deals are quoted in gigawatts, not GPU counts. Google says it cannot build fast enough and is buying third-party capacity.

AVOID LOCK-IN

Standardize on MCP and A2A at the protocol layer and route models at runtime. Four frontier models shipped in three weeks in July.

TCO REALITY

Inference spend now exceeds training spend. Layer 3 decides your gross margin, not Layer 5.

VALUE ACCRUAL

Value keeps moving up the stack. Vertical applications hold their multiples better than horizontal ones.

REGULATORY CLOCK

EU AI Act transparency rules apply August 2, 2026. High-risk rules slipped to December 2027. Start the inventory now anyway.

19 Terms Every Executive Needs

LLM

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

Foundation Model

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

Tokens

Units of text, roughly words or subwords. Usage is priced per token.

Inference

Running a trained model to produce predictions or text.

Fine-Tuning

Customizing a pre-trained model on your domain data.

Context Engineering

Deciding what information a model gets handed at request time.

RAG

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

Vector Database

Stores embeddings for semantic search and retrieval.

Multimodal

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

Neocloud

GPU-specialized cloud providers built purely for AI workloads.

AI Agent

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

Agent Identity

Credentials and permissions issued to a non-human actor. The unsolved problem underneath most agent security incidents.

MCP

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

A2A

Agent-to-Agent. Standard for agents delegating work to each other.

Evals

Automated tests that measure whether model output is actually good.

Guardrails

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

Hallucination

When the model confidently generates false information.

Open Weights

Downloadable model weights, licensed for self-hosting. Not the same as open source.

Model Routing

Sending each request to the cheapest model that can handle it.

This reference guide is designed for executive decision-makers navigating AI strategy.

Companies listed represent significant players as of July 2026. This landscape changes rapidly. Valuations and share figures mix audited filings, company statements and private-company reporting.

(c) 2026 Chris Labatt-Simon