Created by Chris Labatt-Simon | labattsimon.com

EXECUTIVE REFERENCE -- 2026 EDITION

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

PRODUCTIVITY
- MS Copilot
- Google AI
- Notion AI

CODE & DEV
- Claude Code
- Cursor
- GitHub Copilot
- OpenAI Codex
- Replit

AGENT PLATFORMS
- NVIDIA Toolkit
- Wukong

CUSTOMER CX
- Intercom Fin
- Zendesk AI

VERTICAL AI
- Harvey
- Clinomic
- Abridge
- Bloomberg

CONTENT & MEDIA
- Seedance
- Midjourney
- ElevenLabs
- ~~Sora~~
- Runway
- DALL-E 3
- Stable Diffusion
- Suno
- Pika
- Kling
- PlayHT
- Ideogram
- Flux

LAYER 07 — Safety & Governance

Red-teaming, guardrails, compliance & risk manag

- Protect AI
- Robust Intelligence
- Arthur AI
- Guardrails AI
- CalypsoAI
- Check Point
- Credo AI
- HiddenLayer
- Patronus AI

LAYER 06 — Orchestration & Agents

Frameworks, agent toolkits & workflow engines

- LangChain
- OpenClaw
- CrewAI
- LlamaIndex
- OpenAI Agents SDK
- Google ADK
- NemoClaw
- AutoGen
- Semantic Kernel
- Haystack
- Vercel AI SDK
- Anthropic MCP
- Dify
- n8n
- Galileo Agent Ctrl

LAYER 05 — Model Dev & MLOps

Training, fine-tuning, evaluation & deployment

- Hugging Face
- Weights & Biases
- Scale AI
- Together AI
- Replicate
- Anyscale
- Labelbox
- MLflow
- Modal
- Fireworks AI
- BentoML
- Ollama
- SiliconFlow

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)

LAYER 03 — Data & Vector Storage

Databases, vector stores, data pipelines & ETL

- Snowflake
- Databricks
- Pinecone
- Weaviate
- Chroma
- MongoDB Atlas
- pgvector
- Elastic
- Qdrant
- Milvus
- Fivetran
- Unstructured

LAYER 02 — Cloud & Compute

Cloud platforms, GPU clusters & inference hosting

- AWS
- Microsoft Azure
- Google Cloud
- CoreWeave
- Lambda Labs
- Oracle Cloud
- IBM Cloud
- Vultr
- Crusoe Energy
- Hetzner

LAYER 01 — Silicon & Hardware

GPUs, TPUs, custom silicon & accelerators

- NVIDIA
- AMD
- Intel
- Google (TPUs)
- Cerebras
- Groq
- SambaNova
- Qualcomm
- Apple Silicon
- d-Matrix

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.

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

Companies listed represent significant players as of March 2026 -- this landscape changes rapidly.

LAYER 04

Foundation Models

Large language, vision & multimodal models