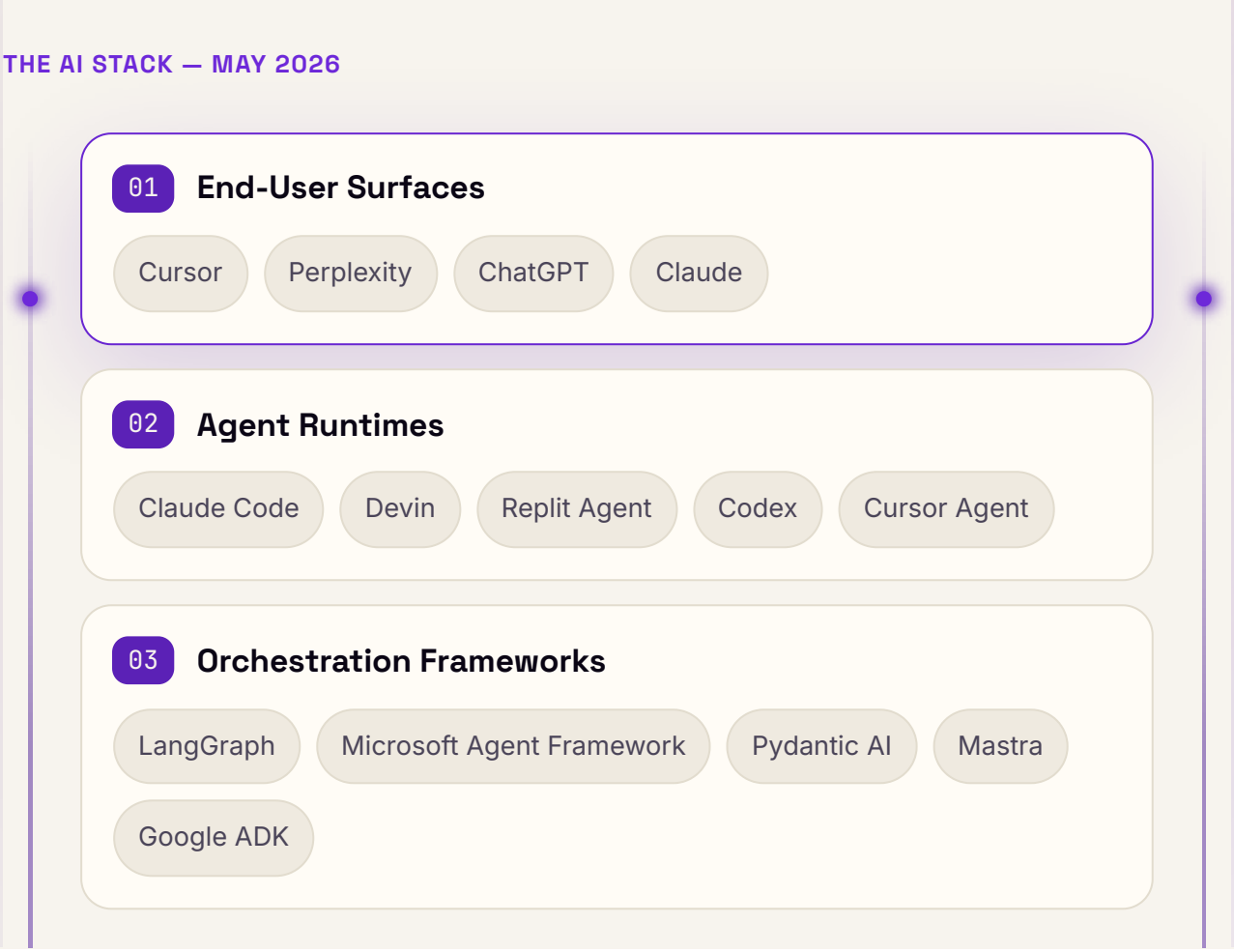
> kb/ai-stack-2026.md · STACK · 12 MIN

FIELD MAP

THE AI STACK — MAY 2026

The application layer most teams ship on is now ten distinct layers deep, wrapped by two rails that touch every one of them. This is a working map, not a buyer's guide: where each category sits, what a few representative providers do, and how the pieces connect.

Read it top to bottom — surface to silicon. The left rail, Observability, and the right rail, Governance, are not steps in the flow; they are concerns that cut across all ten layers. Tap any provider in the diagram to jump to its explanation and an outbound link below.



OBSERVABILITY

GOVERNANCE

Tap any item for details ↓

01 END-USER SURFACES

Cursor

AI-first code editor; agentic edits and codebase-wide changes from natural language.

Visit provider ↗

Perplexity

Answer engine: conversational search with live sources and citations.

Visit provider ↗

ChatGPT

OpenAI's consumer assistant for chat, reasoning and tool use.

Visit provider ↗

Claude

Anthropic's assistant across web, desktop and mobile, tuned for long-context work.

Visit provider ↗

02 AGENT RUNTIMES

Claude Code

Terminal-native agentic coding from Anthropic; delegates multi-step engineering tasks.

Visit provider ↗

Devin

Cognition's autonomous software engineer that plans and executes end-to-end.

Visit provider ↗

Replit Agent

Builds and deploys full apps from a prompt inside Replit's cloud IDE.

Visit provider ↗

Codex

OpenAI's coding agent for the cloud and CLI, running tasks in isolated sandboxes.

Visit provider ↗

Cursor Agent

Cursor's background agent mode for parallel, longer-running coding work.

Visit provider ↗

03 ORCHESTRATION FRAMEWORKS

LangGraph

Graph-based orchestration for stateful, multi-step agent workflows (LangChain).

Visit provider ↗

Microsoft Agent Framework

Microsoft's unified agent framework, consolidating Semantic Kernel and AutoGen.

Visit provider ↗

Pydantic AI

Type-safe Python agent framework from the Pydantic team.

Visit provider ↗

Mastra

TypeScript framework bundling agents, workflows, memory and evals.

Visit provider ↗

Google ADK

Google's open-source Agent Development Kit (Python, Java, Go, TypeScript).

Visit provider ↗

04 PROTOCOL LAYER

MCP

Model Context Protocol (Anthropic): a standard way to connect models to tools and data.

A2A

Agent2Agent: cross-vendor agent interoperability; created by Google, now Linux Foundation.

Visit provider ↗

Visit provider ↗

AG-UI

Agent-User Interaction protocol (CopilotKit): event stream between agent backends and frontends.

Visit provider ↗

05 MEMORY

Mem0

Drop-in memory API combining vector, graph and key-value stores for personalization.

Visit provider ↗

Letta

OS-style agent memory with paging between context and archival storage (formerly MemGPT).

Visit provider ↗

Zep

Temporal knowledge-graph memory (Graphiti) that tracks how facts change over time.

Visit provider ↗

06 RETRIEVAL

Cohere Rerank

Reranking models that reorder candidate passages by true relevance.

Voyage AI

High-quality embedding and reranking models (part of MongoDB).

Visit provider ↗

Visit provider ↗

Neo4j GraphRAG

Graph-based RAG that grounds retrieval in a knowledge graph.

Visit provider ↗

Elastic

Hybrid keyword and vector search on the Elasticsearch engine.

Visit provider ↗

07 STORAGE

pgvector

Postgres extension adding vector similarity search to an existing database.

Visit provider ↗

Qdrant

Open-source vector database with payload filtering and hybrid search.

Visit provider ↗

Turbopuffer

Serverless vector and full-text search built on object storage for low cost at scale.

Visit provider ↗

Pinecone

Fully managed vector database for production retrieval.

Visit provider ↗

neo4j

Native graph database for richly connected data.

Visit provider ↗

08 MODEL GATEWAY

Portkey

AI gateway adding routing, caching, guardrails and observability across providers.

Visit provider ↗

LiteLLM

Unified SDK and proxy exposing 100+ model providers behind one OpenAI-style API.

Visit provider ↗

OpenRouter

A single API that routes requests across many models and providers.

Visit provider ↗

09 FOUNDATION MODELS

Claude (Anthropic)

Anthropic's Claude model family, tuned for reasoning, coding and long context.

Visit provider ↗

GPT (OpenAI)

OpenAI's GPT family of general-purpose frontier models.

Visit provider ↗

Gemini (Google)

Google DeepMind's multimodal Gemini model family.

Visit provider ↗

Meta (Llama)

Meta's open-weight Llama models for self-hosting and fine-tuning.

Visit provider ↗

DeepSeek

Open-weight models known for strong reasoning at low cost.

Visit provider ↗

Qwen

Alibaba's open-weight Qwen model family across sizes and modalities.

Visit provider ↗

10 INFERENCE + COMPUTE

Together AI

Inference cloud for running and fine-tuning open models at scale.

Visit provider ↗

Fireworks AI

Fast, cost-efficient inference serving for open models.

Visit provider ↗

vLLM

Open-source high-throughput inference engine for LLM serving.

Visit provider ↗

NVIDIA

Data-center GPUs that dominate AI training and inference.

Visit provider ↗

AMD MI400

AMD's Instinct MI400-series AI accelerators — AMD's datacenter challenge to NVIDIA.

Visit provider ↗

Google TPU

Google's Tensor Processing Units for training and serving on Google Cloud.

Visit provider ↗

AWS

Cloud plus custom Trainium and Inferentia silicon for AI workloads.

Visit provider ↗

Groq

LPU-based inference delivering very low-latency token generation.

Visit provider ↗