AI Stack

The AI stack combines integrated tools, libraries, and solutions to create applications with generative AI capabilities.

Programming Language

Model Provider

pretrained large models you can call or fine-tune

Category	Providers	Notes
General-purpose LLMs	OpenAI (GPT-4/5), Anthropic (Claude 3), Google DeepMind (Gemini 1.5), Mistral (Mixtral), Cohere, xAI (Grok)	Provide reasoning, coding, and multimodal capabilities via APIs.
Open-source LLMs	Meta (Llama 3/3.1), Mistral, Falcon 2, Yi, Command R+ (Cohere)	Self-hostable; widely used for private deployments.
Specialized models	Hugging Face Hub, Replicate, Anthropic Claude Haiku, Stability AI, Runway	Domain-specific or multimodal (text-image-video).

LLM Orchestrators & FW

how models interact — prompt chaining, memory, and multi-agent coordination.

Category	Tools/Frameworks	Notes
Prompt orchestration / agent frameworks	**LangChain**, **LlamaIndex**, **Haystack**, **Semantic Kernel (Microsoft)**, **CrewAI**, **AutoGen (Microsoft Research)**	Manage complex workflows and context across multiple LLM calls.
Multi-agent systems	**OpenDevin**, **SuperAGI**, **ChatDev**, **AutoGPT**, **MetaGPT**, **LangGraph**	Enable coordination between multiple AI agents for goal-based execution.
Low-code / no-code AI builders	**Dust**, **Flowise**, **Pinecone Canopy**, **Relevance AI**, **PromptLayer**	Visual or declarative orchestration for non-engineers.

Operational and Vector search

retrieval-augmented generation (RAG), **memory**, and **context management**.

Type	Tools	Notes
Vector databases	**Pinecone**, **Weaviate**, **Milvus**, **Qdrant**, **Chroma**, **FAISS (Meta)**	Store embeddings for semantic search & memory.
Document & metadata stores	**MongoDB**, **PostgreSQL (pgvector)**, **Elasticsearch**, **Neo4j (GraphDB)**	Used for hybrid retrieval (structured + semantic).
Operational data layers	**Redis**, **DuckDB**, **SQLite**, **Delta Lake (Databricks)**	Cache, ephemeral state, and structured storage for agent contexts.

Monitoring and Observability

Track model performance, drift, latency, hallucination rates, and cost metrics.

Category	Tools	Notes
LLM analytics / tracing	**LangFuse**, **Helicone**, **PromptLayer**, **Weights & Biases (W&B)**, **Arize AI**, **Traceloop**, **Phoenix (Arize)**	Capture traces, cost, latency, token usage, and output quality.
Evaluation frameworks	**TruLens**, **Evals (OpenAI)**, **Ragas**, **Giskard**, **DeepEval**	Evaluate response correctness and robustness systematically.
Ops observability	**Grafana**, **Prometheus**, **Datadog**, **New Relic**	For system-level metrics and uptime monitoring.

Deployment

Where models or agent systems are hosted and scaled.

Category	Tools / Platforms	Notes
Model serving	**vLLM**, **Ollama**, **Triton Inference Server (NVIDIA)**, **Ray Serve**, **TGI (Text Generation Inference)**	Efficient inference servers for LLMs.
Cloud deployment	**Azure OpenAI**, **AWS Bedrock**, **Google Vertex AI**, **Databricks Mosaic AI**, **Modal**, **RunPod**	Managed infrastructure for hosting and scaling AI workloads.
Container / on-prem	**Docker**, **Kubernetes**, **Hugging Face Inference Endpoints**, **Replicate**, **Banana.dev**	Private or hybrid deployments for enterprise compliance.
Edge / local inference	**LM Studio**, **Ollama**, **Llama.cpp**, **Jan.ai**	Run models directly on consumer hardware or private servers.

AI Stack

The AI stack combines integrated tools, libraries, and solutions to create applications with generative AI capabilities.

Programming Language

Model Provider

pretrained large models you can call or fine-tune

LLM Orchestrators & FW

how models interact — prompt chaining, memory, and multi-agent coordination.

Operational and Vector search

retrieval-augmented generation (RAG), **memory**, and **context management**.

Monitoring and Observability

Track model performance, drift, latency, hallucination rates, and cost metrics.

Deployment

Where models or agent systems are hosted and scaled.

Category	Tools / Platforms	Notes
Model serving	vLLM, Ollama, Triton Inference Server (NVIDIA), Ray Serve, TGI (Text Generation Inference)	Efficient inference servers for LLMs.
Cloud deployment	Azure OpenAI, AWS Bedrock, Google Vertex AI, Databricks Mosaic AI, Modal, RunPod	Managed infrastructure for hosting and scaling AI workloads.
Container / on-prem	Docker, Kubernetes, Hugging Face Inference Endpoints, Replicate, Banana.dev	Private or hybrid deployments for enterprise compliance.
Edge / local inference	LM Studio, Ollama, Llama.cpp, Jan.ai	Run models directly on consumer hardware or private servers.

Information Retrieval Mechanisms

What are Information Retrieval Mechanisms?