AI Tech Stack 2025 for Solution Architects

Core Model Layer

- LLMs:
  Tools: OpenAI GPT-4-turbo, Claude 3, Gemini 1.5, Mistral, Mixtral, LLaMA 3
  Purpose: Hosted and open-source models for various use cases
- Embeddings:
  Tools: OpenAI Ada v3, BGE, Cohere, Hugging Face
  Purpose: Embed documents for semantic search and RAG

RAG Layer

- Vector Store:
  Tools: ChromaDB, Qdrant, Pinecone, FAISS, Weaviate
  Purpose: Store and retrieve vector embeddings
- RAG Framework:
  Tools: LangChain, LlamaIndex, Haystack
  Purpose: Build RAG pipelines
- Document Loaders:
  Tools: LangChain Loaders, Unstructured.io, PyMuPDF
  Purpose: Convert and chunk source documents
- Chunking Strategy:
  Tools: Recursive splitter, Metadata-aware
  Purpose: Improve RAG performance

Agentic Layer

- Agent Frameworks:
  Tools: CrewAI, AutoGen, LangGraph, OpenAgents
  Purpose: Multi-agent orchestration
- Agent Memory:
  Tools: LangChain Memory, AutoGen Custom Memory
  Purpose: Maintain context between tasks
- Tool Integration:
  Tools: APIs, DB tools, browser tools
  Purpose: Agents interact with real-world sources

Backend & API Layer

- API Framework:
  Tools: FastAPI, Flask, Express.js
  Purpose: Backend APIs for LLM apps
- Orchestration:

Tools: LangGraph, Prefect, Airflow

Purpose: Handle task and logic flows

- Auth & Rate Limit:

Tools: OAuth2, Auth0, Kong

Purpose: Secure and protect API services

Frontend Layer

- UI Framework:

Tools: React + Tailwind, Next.js, Streamlit, Gradio

Purpose: Create user interface

- Chat UI:

Tools: react-chat-widget, botpress

Purpose: Integrate LLM chat UI

Deployment & Ops

- Containers:

Tools: Docker

Purpose: Package services in containers

- Infra & Cloud:

Tools: AWS Bedrock, Azure, GCP, Railway

Purpose: Serve and host AI apps

- Monitoring:

Tools: LangFuse, Arize AI, Grafana

Purpose: Monitor LLM usage & performance

- LLMOps:

Tools: W&B, BentoML, MLflow

Purpose: Track, version, and serve models

Testing & Evaluation

- Prompt Evaluation:

Tools: DeepEval, TruLens, Promptfoo

Purpose: Test for hallucinations, quality

- Unit Testing:

Tools: pytest, LangChain testing

Purpose: Verify app logic

Learning Resources

- LangChain Docs:

Tools: https://docs.langchain.com

Purpose: Documentation for LangChain

- LlamaIndex Docs:

Tools: https://docs.llamaindex.ai

Purpose: Docs for LlamaIndex framework

- AutoGen:

Tools: https://microsoft.github.io/autogen

Purpose: Multi-agent framework by Microsoft

- CrewAI:

Tools: https://docs.crewai.com

Purpose: Docs for CrewAI agent framework

- OpenAI API Docs:

Tools: https://platform.openai.com/docs

Purpose: Official OpenAI documentation

- Vector DB Guide:

Tools: https://www.pinecone.io/learn

Purpose: In-depth RAG and vector DB guide