What does a real Agentic AI stack look like?  
  
This is the blueprint.  
  
The messy experimentation phase of AI agents is giving way to real engineering architectures.  
  
Here’s a breakdown of the full Agentic AI infrastructure stack used by cutting-edge teams 👇  
  
1️⃣ Set the Goal → AI Agent Orchestration  
You define objectives, manage workflows, and delegate tasks.  
Popular frameworks:  
→ LangGraph, [**AutoGen**](https://www.linkedin.com/company/pyautogen/), [**CrewAI**](https://www.linkedin.com/company/crewai-inc/)  
→ [**OpenAI**](https://www.linkedin.com/company/openai/) functions, [**LlamaIndex**](https://www.linkedin.com/company/llamaindex/), [**LangChain**](https://www.linkedin.com/company/langchain/)  
→ DSPy, Semantic Kernel, Rasa, AGiXT  
  
💡 This layer enables reasoning, planning, and coordination across tools and models.  
  
2️⃣ Take Action → Tool Calls + Caching  
Agents interact with external APIs or tools to accomplish tasks.  
→ Tool Calls: [**Zapier**](https://www.linkedin.com/company/zapier/), ChartGPT, @WolframAlpha  
→ Caching: [**Redis**](https://www.linkedin.com/company/redisinc/), [**Zilliz**](https://www.linkedin.com/company/zilliz/) (commonly used for low-latency retrieval + vector storage)  
  
💡 Caching improves speed and efficiency for repeated queries and state handling.  
  
3️⃣ Acquire Information  
a) AI Models  
OpenAI, [**Anthropic**](https://www.linkedin.com/company/anthropicresearch/), Meta, [**Mistral AI**](https://www.linkedin.com/company/mistralai/), Gemini, [**Cohere**](https://www.linkedin.com/company/cohere-ai/), [**Hugging Face**](https://www.linkedin.com/company/huggingface/)  
  
b) Agent Memory (short-term + long-term)  
MemGPT, LangChain, Letta, Redis, [**Mem0**](https://www.linkedin.com/company/mem0/)  
  
💡 Enables stateful multi-step reasoning and long-horizon tasks.  
  
c) Data Sources  
[**Snowflake**](https://www.linkedin.com/company/snowflake-computing/), [**Databricks**](https://www.linkedin.com/company/databricks/), Redis, Oracle  
  
d) Embedding Models  
OpenAI, Anthropic, Vertex AI, Cohere, Mistral, Hugging Face, Jina AI  
  
e) Vector Databases  
Pinecone, Chroma, PostgreSQL, Redis, [**Milvus, created by Zilliz**](https://www.linkedin.com/company/the-milvus-project/), [**Weaviate**](https://www.linkedin.com/company/weaviate-io/)  
  
💡 Retrieval-Augmented Generation (RAG) is essential for knowledge-grounded agents.  
  
4️⃣ Scale with Infrastructure  
a) Hosting  
AWS, Microsoft Azure, Google Cloud  
  
b) Hardware  
[**NVIDIA**](https://www.linkedin.com/company/nvidia/), Intel, Groq  
  
c) MLOps  
LangSmith, [**Langfuse**](https://www.linkedin.com/company/langfuse/), [**MLflow**](https://www.linkedin.com/company/mlflow-org/), [**Weights & Biases**](https://www.linkedin.com/company/wandb/)  
  
d) Authentication  
Clerk, [**Kong Inc.**](https://www.linkedin.com/company/konghq/)  
  
e) Developer Tools  
[**FlowiseAI (YC S23)**](https://www.linkedin.com/company/flowiseai/)  
  
💡 This layer transforms prototypes into production-ready, enterprise-grade agent systems.  
  
👉 If you’re designing AI agents for real-world use (not just prototypes), this architecture gives you the reference blueprint.  
  
This is the Agentic AI infrastructure map of 2025.  
  
Massive credit to the open-source ecosystem and the emerging tools making it possible.  
  
Follow [**Pallavi**](https://www.linkedin.com/in/progressivethinker/) for more such insights.

Activate to view larger image,

