Tech Stack for Agentic AI

What most think of AI Agents is just the tip of the iceberg. Underneath the infrastructure is a completely different story. Unlike most traditional automation, which relies on conventional ML practices. Every component of AI Agents is modular and industry-standard.

These are very specialized startups working on solving specific problems with AI agents and improving them for scalable automation.
Let's briefly learn about those providers:

- CPU/GPU Providers

➤ These power AI agents with the necessary computing resources. They handle large-scale model training, inference, and optimization while ensuring security and cost efficiency.

- Infra/Base

➤ This layer ensures containerization, deployment, and connectivity for AI agents. It allows for scalability, efficient resource allocation, and stable infrastructure to run complex AI applications.

- Database

➤ AI agents require structured and unstructured data to function effectively. Databases in this stack store, retrieve, and manage data, enabling agents to maintain context, learn over time, and optimize responses.

- Foundational Models

➤ These serve as the "brains" of AI agents. Large language models (LLMs), retrieval-augmented models (LRMs), and small language models (SLMs) power various agentic use cases, from text generation to decision-making.

- Model Routing

➤ A crucial part of the stack, it directs user queries to the most suitable AI model based on task complexity, cost, or latency, ensuring optimal performance and efficiency.

- Agent Orchestration

➤It enables automation, multi-agent collaboration, and structured decision-making in real-world applications.

- Agentic Observability

➤ This layer provides visibility into agent behavior, metrics, prompt logs, and real-time analytics.

- Tools

➤ AI agents often rely on external APIs, search engines, and third-party integrations to enhance their capabilities. These tools help agents retrieve real-time information, automate workflows, and extend functionality.

- Authentication

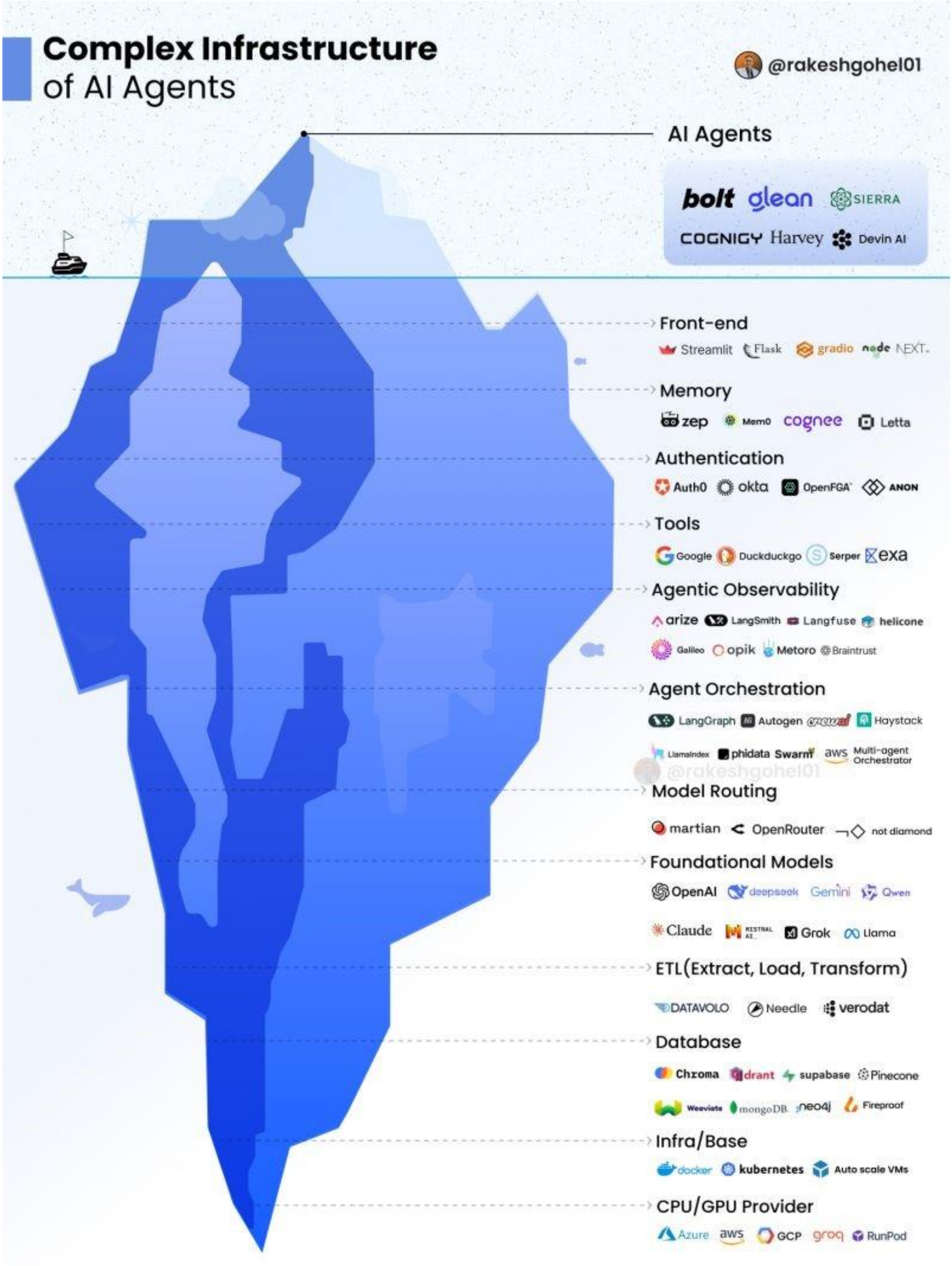
➤ Security and access control are critical when deploying AI agents. This layer ensures secure identity verification, access management, and data privacy compliance.

- Memory

➤ This layer allows agents to retain user interactions, store contextual knowledge, and improve over time.

- Front-end

➤ The user interface is essential for interacting with AI agents. This layer includes web and app frameworks that enable seamless communication between users and agents.



Complex Infrastructure of AI Agents

@rakeshgohel01

AI Agents

bolt glean SIERRA
COGNIGY Harvey Devin AI

Front-end

Streamlit Flask gradio node NEXT.

Memory

zep Mem0 cognee Letta

Authentication

Auth0 okta OpenFGA ANON

Tools

Google Duckduckgo Serper exa

Agentic Observability

arize LangSmith Langfuse helicone
Galileo opik Metoro Braintrust

Agent Orchestration

LangGraph Autogen crewai Haystack
LlamaIndex phidata Swarm aws Multi-agent Orchestrator

Model Routing

martian OpenRouter not diamond

Foundational Models

OpenAI deepseek Gemini Qwen
Claude MISTRAL AI Grok Llama

ETL(Extract, Load, Transform)

DATAVOLO Needle verodat

Database

Chroma qdrant supabase Pinecone
Weaviate mongoDB neo4j Fireproof

Infra/Base

docker kubernetes Auto scale VMs

CPU/GPU Provider

Azure aws GCP groq RunPod

Image credit Rakesh Gohel