AGENTIC RAG
TECH STACK 2025

A LAYERED BREAKDOWN OF
TOOLS POWERING AI AGENTS

7 ALIGNMENT
6 MEMORY
5 DATA EXTRACTION
4 EMBEDDING
3 VECTOR DB
2 FRAMEWORK
1 LLMs
0 DEPLOYMENT

SLIDE TO EXPLORE →

Why Agentic RAG Matters

🔹 **AI Agents are only as powerful as their infrastructure** : An agent's intelligence doesn't just come from the model — it comes from the ecosystem around it. Without the right stack, even the best LLMs underperform.

🔹 **A strong stack = precise, context-aware retrieval** : Agentic RAG ensures that responses aren't generic. Instead, they're grounded in accurate, real-time data and tailored to the context of the task.

🔹 **Every serious company is chasing its ideal stack:** Enterprises are investing heavily to build robust RAG systems that balance speed, scalability, accuracy, and safety. This stack becomes their AI backbone.

🔹 **Why it matters for the future** : When done right, Agentic RAG transforms LLMs into trustworthy, adaptive agents — capable of continuous learning, reasoning, and decision-making.

👉 **Next: Let's break down how each layer of the stack fits together.**

Layer 0: Deployment

Purpose: The foundation where agents live.

- **Host & run agents on cloud / GPU infrastructure.**
- **Ensure scalability, reliability, and speed.**
- **Tools: Groq, AWS, Google Cloud, Together.ai.**

💡 **Without solid deployment, your agents can't scale.**

together.ai groq

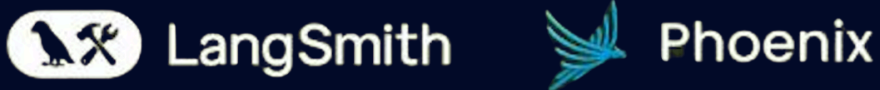
Google Cloud aws

Layer 1: Evaluation

Purpose: Measure and improve retrieval quality.

- Continuous testing & scoring.
- Identify weak points in reasoning.
- Collect feedback to improve performance.
- Tools: LangSmith, Phoenix, DeepEval, Ragas.

💡 “You can’t improve what you don’t measure.”

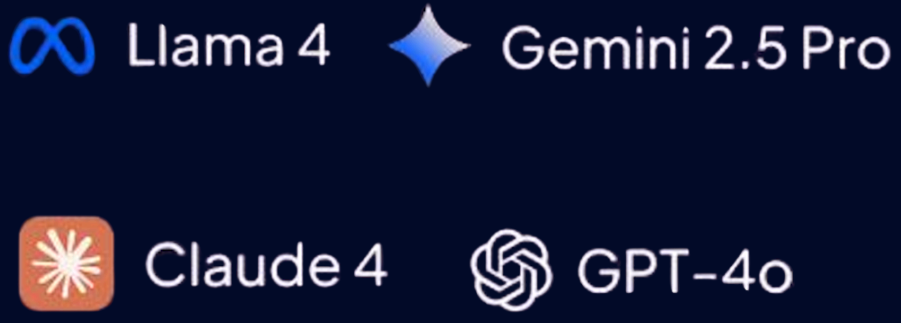


Layer 2: LLMs (The Brain)

Purpose: The reasoning core.

- **Powers conversation, planning, and decision-making.**
- **Large language models handle complex instructions.**
- **Options: Llama 4, Gemini 2.5 Pro, Claude 4, GPT-4o.**

💡 **Think of this as the thinking engine of your system.**

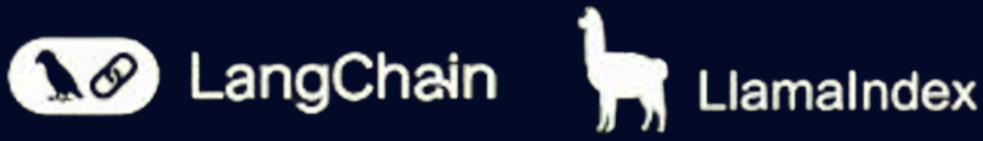


Layer 3: Frameworks

Purpose: The orchestration layer.

- Manage tool usage, memory, and prompts.
- Enable multi-agent workflows.
- Tools: LangChain, LlamaIndex, Haystack, DSPy.

💡 Frameworks = the glue that connects everything.



Layer 4: Vector Databases

Purpose: Retrieval engine.

- Store dense vector embeddings of knowledge.
- Enable similarity search over huge corpora.
- Tools: Pinecone, Chroma, Milvus, Weaviate.

💡 The memory vault for fast, relevant lookups.

Layer 5: Embeddings

Purpose: Convert raw text into math.

- Represent concepts as dense vectors.
- Enable semantic similarity & grounding.
- Tools: Nomic, Ollama, Voyage AI, OpenAI.

💡 Embeddings = the language of meaning.

NOMIC Ollama

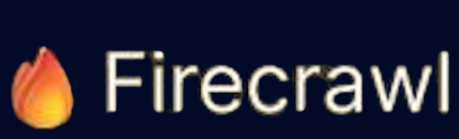
VOYAGE AI OpenAI

Layer 6: Data Extraction

Purpose: Pull fresh data in real time.

- Parse websites, PDFs, and APIs.
- Transform raw info into structured knowledge.
- Tools: Firecrawl, Scrapy, Docling, Llamaparse.

💡 Agents stay smart only if fed with new data.



Layer 7: Memory

Purpose: Give agents persistence.

- Maintain context across sessions.
- Personalize interactions.
- Tools: Zep, Mem0, Cogngee, Letta.

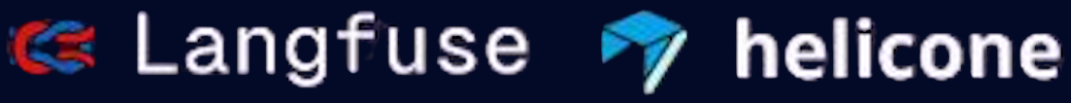
💡 Memory = from one-off chatbots → lifelong AI assistants.

Layer 8: Alignment & Observability

Purpose: Safety & oversight.

- Validate outputs.
- Ensure behavior matches expectations.
- Tools: Guardrails AI, Arize, Langfuse, Helicone.

💡 Without alignment, your agents risk going off-track.



Interested in more content like this?

Follow me :

OM NALINDE