AI Engineer – RAG & LangChain (LLM Systems)

Role Overview

We are looking for a strong AI Engineer who can design, build, and optimize production-grade Retrieval-Augmented Generation (RAG) systems using LangChain and modern LLM infrastructure. This is not a prompt engineering role. We need someone who understands embeddings, vector databases, chunking strategies, context optimization, hallucination control, and cost-performance tradeoffs.

Core Responsibilities

- Design and implement scalable RAG pipelines
- Optimize embedding and chunking strategies
- Build multi-retriever and hybrid search systems
- Develop document ingestion pipelines (PDF, DOCX, Web, DB)
- Implement LangChain agents, tools, and memory systems
- Integrate vector databases with metadata filtering
- Optimize token usage, latency, and cost
- Build evaluation pipelines and hallucination mitigation systems

Required Technical Skills

- Strong Python and backend engineering fundamentals
- LangChain experience (mandatory)
- Production RAG implementation experience
- Vector databases (Pinecone, Weaviate, Qdrant, Milvus, pgvector)
- FastAPI or REST API development
- Docker and basic distributed systems knowledge

Nice to Have

- LangGraph or LlamaIndex
- Kafka, Redis, Kubernetes
- Fine-tuning (LoRA / QLoRA)
- Model quantization
- Graph RAG or knowledge graphs
- Self-hosted LLM inference optimization