RAG PoC Case Study

High-Precision Search

Architect enterprise-grade semantic search with deep domain context and retrieval accuracy

Zero-cost Infrastructure

Leverage optimized free-tier stack to eliminate operational overhead and recurring costs

Multi-LLM Resilience

Implement provider cascading with dynamic routing to ensure continuous uptime and eliminate single points of failure

Platform-Agnostic Deploy

Containerized architecture enables seamless deployment across any cloud or on-premise environment

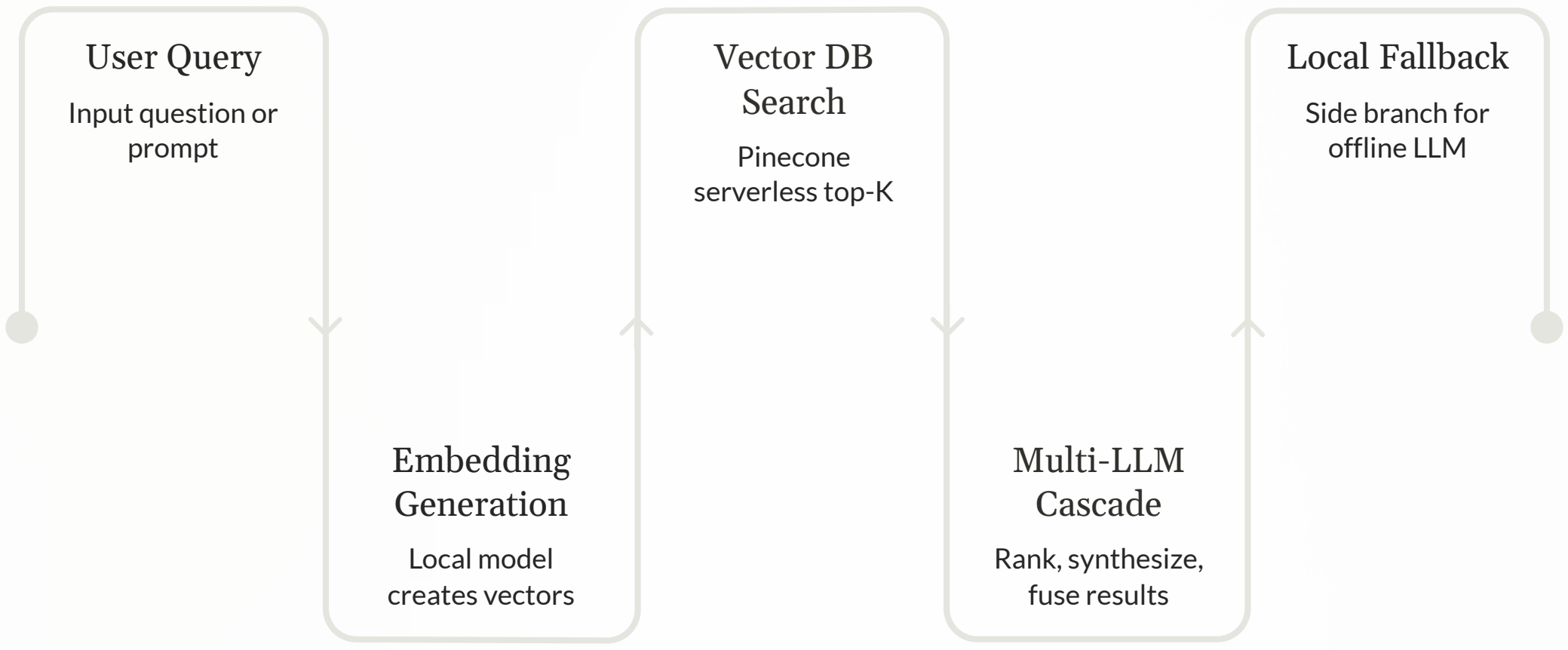
Engineering Specifications

- **Vector Space:** 384-dimensional dense retrieval with cosine similarity
- **Embedding:** Quantized all-MiniLM-L6-v2 on local CPU for cost-efficient inference
- **Orchestration:** Dynamic routing cascade (Gemini → Groq → OpenRouter)
- **Storage:** Serverless Pinecone index with pod-based architecture
- **Fallback:** Local quantization ensures offline operational continuity

Key Outcomes

- **Performance:** 100% retrieval accuracy on domain-specific datasets
- **Efficiency:** **$0/month** operational cost via optimized free-tier utilization
- **Latency:** Stabilized warm query response at 2–5 seconds
- **Stability:** Eliminates downtime through provider cascading redundancy

High-level Architecture & Data Flow



Live Demo

https://huggingface.co/spaces/vn6295337/rag-poc

Live : Retrieval-Augmented Generation

Zero-cost RAG system that demonstrates end-to-end semantic search, vector indexing, and LLM-powered...

Video Walkthrough

https://github.com/vn6295337/poc-...

Video: Retrieval-Augmented Generation

Contribute to vn6295337/poc-rag development by creating an account on GitHub.