Role Overview:

We're looking for a GenAI Developer to design, build, and deploy intelligent LLM-powered systems—from single-agent chatbots, copilots to complex multi-agent applications—at scale. We are particularly interested in candidates who have hands-on experience in taking GenAI applications from concept to production, especially within high-volume B2C environments. This role prioritizes individuals who understand the nuances of deploying, maintaining, and optimizing GenAI solutions for real-world users, beyond the scope of Proof-of-Concept (PoC) development. You will work across the full stack, integrating LLMs, microservices, vector databases, backend APIs, and modern cloud infrastructure.

Key Responsibilities:

1. GenAI Application Development & Deployment

- **Develop scalable, asynchronous microservices using Python (FastAPI) for chatbots, copilots, and agentic workflows.**
- **Design event-driven architectures to support high concurrency, rate limiting, and real-time responsiveness.**
- **Implement secure, versioned REST/gRPC APIs**
- **Use Pydantic, dependency injection, and modular coding practices for maintainability.**
- **Proficient in working with databases using ORMs like SQLAlchemy**
- **Ensure observability using logging, metrics, tracing, and health checks.**
- **Create responsive React.js frontends integrated via REST APIs or WebSockets.**
- **Deploy applications on Cloud Run, GKE, using Docker, Artifact registry, CI/CD pipelines**

2. LLM-Powered Conversational Interfaces

- **Design and build LLM-powered chatbots, voicebots, copilots and other applications using LangChain or custom orchestration frameworks.**
- **Integrate enterprise-grade LLM APIs (Gemini, OpenAI, Claude) for multi-turn, intelligent interactions.**
- **Implement user session management and context/state tracking for personalized and continuous conversations.**
- **Build RAG pipelines with vector databases, knowledge graphs to ground responses with external knowledge and documents.**

- **Apply advanced prompt engineering (ReAct, Chain-of-Thought with tool calling) for precise and goal-oriented outputs.**
- **Ensure performance in low-latency, streaming environments using WebSockets, gRPC, and SIP media gateways.**
- **Perform fine-tuning of open-source LLMs (LLaMA variants) using techniques like SFT, LoRA, for cost-effective domain adaptation.**
- **Optimize high-speed inference pipelines leveraging multi-GPU clusters (up to 8x H100s) to reduce latency and improve throughput.**

3. Multi-Agent Systems & Orchestration

- **Create multi-agent systems & Implement orchestration patterns like supervisor-agent, hierarchical, and networked agents using frameworks like ADK, Pydantic AI and LangGraph.**
- **Use LangGraph for stateful workflows with memory, conditional branching, retries, and async execution.**
- **Enable persistent context and long-term memory**
- **Monitor behavior, drift, and performance using observability tools.**
- **Skilled in developing agents with ADK and A2A protocols & experienced in configuring custom and remote MCP servers.**

Preferred Tech Stack:

- **Languages/Frameworks: Python, FastAPI, HTML, CSS, React.js, LangChain, LangGraph, Pydatic AI, ADK (Agent Development Kit)**
- **LLMs & Agents: OpenAI (GPT-4), Claude, Gemini, Mistral, LLaMA 3.2/4**
- **Databases: BigQuery, Redis, FAISS, Pinecone, SQLAlchemy, Chroma, GCP Vector search**
- **Protocols/APIs: REST, gRPC, WebSockets, OAuth2, OpenAPI, MCP, A2A**

Additional Good to have Tech Stack:

- **DevOps: Docker, GitHub Actions, Jenkins, GKE, Cloud Run**
- **Infra & Tools: GCP, Azure, Pub/Sub,Artifact Registry, NGINX, Langfuse, Postman, Pytest**