Bonsen AI

Generative AI Engineer

About the job

About Us

At **BonsenAI**, we redefine AI innovation for enterprise customers across the US, Europe, Australia, and New Zealand. As pioneers in secure and scalable AI workflows, we deliver impactful AI solutions rapidly, combining cutting-edge technology with unparalleled efficiency.

We're assembling an **elite engineering team** that leverages AI to transform ideas into production-ready software at unprecedented speed. Join us in shaping the future of AI with meaningful solutions for global enterprises.

Key Responsibilities

- **AI Application Development**: Design and implement end-to-end AI solutions using **OpenAI-Python**, **LangChain**, LiteLLM, and custom frameworks. Build production-grade applications integrating RAG systems, chatbots, and text summarization while optimizing for quality, latency and cost using techniques like model quantization and caching.
- **LLM Integration & API Development:** Develop robust APIs and microservices using FastAPI to integrate multiple LLM providers (**Azure OpenAI**, Anthropic, Mistral). Create scalable vector search solutions using **Azure AI Search**/Pinecone/PgVector, implement streaming responses, and build reliable fallback mechanisms for production environments.
- **AI Solution Architecture**: Design architectures for complex AI applications combining LLMs, embedding models, and vector stores. Implement advanced features like semantic caching, hybrid search, and custom routing logic while ensuring security, reliability, and cost optimization.
- **MLOps & Infrastructure**: Deploy and monitor AI applications using Docker, Kubernetes, in Azure cloud services. Build automated CI/CD pipelines for model deployment, implement A/B testing frameworks, and develop observability solutions for tracking model quality, performance and costs.

Required Skills and Qualifications

- **LLM Expertise**: Hands-on experience with prompt engineering, fine-tuning, and deploying LLM-driven applications (Chatbots, Summarization, RAG systems, Structured Data Extraction).
- **RAG & Vector Databases:** Proficient in RAG implementations and vector database operations (e.g., Azure AI Search, Pinecone, FAISS).
- **Azure AI Services**: Strong understanding of Azure OpenAI and Azure AI Search.
- **MLOps & CI/CD**: Knowledge of MLOps practices for managing AI workflows and deploying scalable pipelines.
- **Database Proficiency**: Familiarity with SQL and NoSQL databases, including Azure Cosmos DB.
- **Libraries & Frameworks**: Experience with OpenAI-Python, LiteLLM, Pydentic and related frameworks.

Nice-to-Have

- **Azure Certifications** such as Microsoft Certified: **Azure AI Engineer Associate** or **Azure Solutions Architect Expert**.
- Experience designing and integrating **AI-driven enterprise workflows** within the Microsoft ecosystem, including **Microsoft Sharepoint, Graph API** and **Microsoft Teams**.
- Experience with advanced techniques (reinforcement learning, multi-modal AI).
- Contributions to open-source AI projects.

Soft Skills

- Foster open communication within the team to enhance collaboration and trust.
- Demonstrate commitment to teamwork by delivering high-quality results consistently.
- Approach challenges with an entrepreneurial mindset to drive innovation and efficiency.
- Strategic thinking and problem-solving capabilities.
- Adaptability to a fast-paced, innovation-driven environment.

Why Join Us?

- **Own** client solutions from concept to ROI measurement.
- Be a key player in delivering reliable AI-powered enterprise solutions.
- Work with cutting-edge technology in a collaborative, innovation-driven environment.
- Flexible work environment with hybrid options.
- Competitive salary and opportunities for career advancement.
- BonsenAI is an equal opportunity employer. We embrace diversity and foster an inclusive workplace where everyone thrives.