G Vibu Vignesh

AI Engineer

Coimbatore , Tamil Nadu , IN

gviboognesh99@gmail.com | 9123503384 | GitHub | LinkedIn

About

An accomplished AI Engineer with a strong background in developing high-performance AI inference servers using Python FastAPI and Robyn framework. Proficient in leveraging and integrating cutting-edge AI technologies, including OpenAI, Anthropic, and Hugging Face models, with expertise in fine-tuning Hugging Face models and deploying open-source LLMs via vLLM. Experienced in building robust LLM agents with LangGraph and N8N for automation, and skilled in utilizing Rust's Candle and Candle-vLLM libraries for efficient LLM inference.

Experience

GenAI Developer *DoaZ (Remote) | 2024-04 - Present*

- Improved information retrieval accuracy from 60% to 80% by implementing Contextual Chunking for document preprocessing.
- Utilized DSPy for advanced prompt optimization, enhancing chatbot response quality and relevance.
- Deployed open-source LLMs (Qwen 148 & 32B) on Vastai GPU servers using VLLM for efficient inference.
- Built and maintained FastAPI endpoints for seamless interaction between the AI backend and the frontend application.
- Deployed LLMs using candle-rust library for faster LLM inference using Rust Programming Language

Skills

Python | RAG Applications | LLM Agents | Rust

RAG Application Developer *Upwork (Remote) | 2023-11 - 2024-04*

- Engineered a PDF chatbot featuring a unique user-driven document selection architecture prior to Q&A.
- Developed a CSV-based chatbot leveraging Langchain's Pandas agent framework for data interaction and analysis.
- Created a web-search-enabled chatbot integrating DuckDuckGo to answer queries with real-time information.
- Enhanced overall chatbot functionality using Python and various OpenAI models.

Skills

Python | RAG Applications

Data Scientist Intern *Gilbert Research Institute (Coimbatore) | 2023-08 - 2023-11*

- Developed a medical image analysis solution for Lung and Liver Tumor segmentation.
- Utilized PyTorch and Convolutional Neural Networks (CNNs) to build and train the segmentation models.
- Communicated effectively with faculty and staff, incorporating feedback for improvement.
- Collaborated within teams to analyze problems and develop data-driven solutions.

Skills

Python | Deep Learning

Projects

Async Rust Port Scanner

A simple and fast multithreaded port scanner written in Rust, leveraging Tokio for asynchronous operations and Clap for command-line argument parsing. It efficiently scans a range of TCP ports on a target IP address.

- Built with Rust, ensuring high performance and memory safety.
- Uses Tokio's asynchronous runtime to perform concurrent connection attempts, drastically reducing scan time.
- Features a configurable command-line interface with clap for specifying target IP and custom port ranges.
- Incorporates a timeout mechanism to prevent the program from hanging on unresponsive ports.

Skills

Rust | CLI Tools

Rust Zero-Knowledge Proof Authentication System

A secure client-server authentication system that utilizes Zero-Knowledge Proofs (ZKP) to allow users to authenticate without ever revealing their passwords. The system is built with Rust and uses gRPC for efficient and robust communication.

- Implements Zero-Knowledge Proofs for a secure, privacy-preserving authentication process.
- Uses gRPC with Protocol Buffers for defining service contracts and enabling high-performance, language-agnostic communication.
- Written in Rust, guaranteeing memory safety, concurrency, and high performance.
- Eliminates the risk of password interception by never transmitting the actual password over the network.

Skills

Rust | gRPC | Zero-Knowledge Proofs | Docker

Rust Distributed Chat Server

A distributed server where users can connect and transmit messages to other users

- Distributed Systems in Rust

Skills

TCP

Axum RBAC Workspace System

An open-source backend starter kit for building multi-tenant, workspace-oriented applications using Rust and the Axum framework.

- Built with Rust and Axum for high performance and safety.
- Features secure user authentication, workspace management, and Role-Based Access Control (RBAC).
- Includes database migrations, email notifications, and a modular architecture.

Skills

Axum | Postgresql | sqlx

LangGraph Agent for Candidate Information Extraction

A LangGraph-based conversational agent that interacts with users to extract essential candidate information like contact details, job title, and salary expectations.

- Uses a stateful graph to guide the conversation and ensure all required data is collected.
- Extracts structured data (JSON) from unstructured, natural language user inputs.
- Built with Python, LangGraph, and the OpenAI API.

Open Source Contributions

gpt-researcher

Bug Fix Contribution in gpt-reasearcher project

Skills

Python | LLM Agents

Skills

Python | RAG Applications | LLM Agents | Rust | Deep Learning | CLI Tools | gRPC | Zero-Knowledge Proofs | Docker | TCP | Axum | Postgresql | sqlx

---

Education

Integrated B.Sc. & M.Sc.

Indian Institute Of Technology Madras

Languages

English | Tamil | Hindi