Sai Kumar Kanthala

☎ +919121758794 linkedin.com ✉ saikumar.kanthala143@gmail.com

Education

National Institute of Technology Warangal May 2023
Bachelor of Technology *GPA: 7.48*

Jawahar Navodaya Vidyalaya Bundi,Rajasthan March 2019
PCM *83%*

Skills

Core Concepts:Data Structures & Algorithms(DSA) - C++, OOPS (Object-Oriented Programming Systems)- Java, DBMS (Database Management Systems) - SQL
Languages:Python, Java, SQL, C++, JavaScript
Tools: Git/GitHub, Docker, Langchain, HuggingFace, AWS, Azure
Frameworks: Django,FastAPI,Node.js, Express.js
AI Skills: Retrieval Augmented Generation(RAG),LLM Fine-Tuning, Quantization and Deployment
Databases: PostgreSQL, MySQL, Vector databases - Chroma, Milvus, PgVector

Experience

CloudAngles | *Software Engineer* Sep 2023 – Present

RAG Applications & Development:

- Designed and developed RAG Chatbots that integrate seamlessly with vector databases and prominent data sources like Confluence, SharePoint, and Dropbox.
- Built and deployed Multimodal RAG applications for efficient question-answering over images, text, and structured data.
- Played a key role in the RAG POC development for an enterprise client, showcasing how RAG can automate knowledge retrieval and enhance productivity.
- Responsible for the end-to-end design and low-level architecture of APIs and databases for all RAG-related projects.

MedCodeX – Healthcare Application

- Designed and developed MedCodeX, a healthcare application for medical coders, aimed at improving work efficiency through automated ICD coding suggestions.
- Built REST APIs and designed efficient data models in Django to support seamless data extraction, analysis, and reporting.
- Used LLMs to process medical summaries and generate relevant ICD codes for faster billing and improved accuracy.

LLM Optimization & Deployment

- Implemented GGUF and AWQ quantization techniques for Large Language Models (LLMs), reducing infrastructure costs by 75%.
- Deployed open-source LLMs for Retrieval-Augmented Generation (RAG) applications in AWS and Azure environments.
- Fine-tuned LLMs using PEFT and LoRA techniques to enhance model performance for domain-specific tasks.

TestingAIde - Test Automation Project

- Built a test automation tool that takes in requirements and Jira user stories to generate test cases.
- Enabled users to generate test scripts for each test case, with additional functionality for tagging test cases.
- Developed a chatbot capable of querying the database and providing information using natural language queries.
- Responsible for designing the low-level architecture, APIs, and database for the test automation app, ensuring a scalable and robust solution.

Coursework & Certifications

- The Complete Web Debelopment BootCamp (Udemy)
- Object Oriented Programming & Computer Programming (NIT W CSE Department)