SAI MOHAN REDDY

Hyderabad, Telangana

📞 9063308182 ✉ ssmohanreddy@gmail.com in mohan ⌂ mohan

Summary

Software engineer with 2.5+ years of experience building scalable RESTful APIs using Flask and FastAPI. Skilled in integrating Large Language Models (LLMs) and AWS services to deliver robust back-end systems. Committed to writing clean, well-tested code and building real-world AI-powered applications.

Technical Skills

Languages: Python, HTML/CSS, JavaScript, SQL
Databases: MySQL, PostgreSQL, SQLite, DynamoDB, Milvus
Technologies/Frameworks: Flask, FastAPI, AWS, Large Language Models (LLMs)
Tools: Git, Docker, Postman, AWS CLI, pytest

Experience

Annalect   **Nov 2022 – Current**
Senior Associate - Python Developer   *Hyderabad, Telangana*

- **Developed Gen AI microservice** leveraging asynchronous programming for efficient and scalable management of embedding metadata and approvals, streamlining the workflow for AI-generated embeddings.
- Built a robust **test suite** (unit + integration) using **pytest**, ensuring **90%+ test coverage** and reliable deployment.
- Designed robust **APIs** for interacting with **vector databases**, enabling seamless data retrieval and efficient management of collections.
- Integrated the **Azure OpenAI text-embedding model** into the **Gen AI microservice**, enabling efficient conversion of text data into embeddings.
- Integrated **GCP speech to text API** into microservice for conversion of audio data into text.
- Developed APIs, for ingesting test data from diverse sources(S3, plain text, URL's), making it easier to convert and store the embeddings effectively
- Utilized **Docker** to containerize a microservice, simplifying deployment processes.
- Developed **custom middleware** for **Flask** and **FastAPI** applications centralizing I/O logging to database
- Developed **back-end API (OpenPyxl)** for Excel data exchange. Enables import (new data included), modification in sheets, and reflects changes back to database on import.
- **Implemented custom templates** for data export using OpenPyxl. Clients can modify templates, which are saved to S3 for future use, enabling personalized data exports.

Projects

Classification of Ventricular Arrhythmia | *Python, Numpy, Matplotlib, CNN*   **July 2022**

- Developed a deep learning model to classify **ventricular arrhythmia** based on **ECG** signals.
- Built and trained a **CNN-LSTM** hybrid architecture for accurate classification of arrhythmia patterns
- Preprocessed ECG data using **Pandas** to optimize the input for model training, enhancing diagnosis efficiency.

Multimodal Chat Application | *Python, FastAPI, OpenAI GPT-4, AWS Bedrock, AWS S3*   **March 2024**

- Developed a chat application supporting both **text and image inputs**, leveraging **OpenAI GPT-4** and **AWS Bedrock** for multimodal responses.
- Built **FastAPI**-based REST APIs to handle user inputs, with **AWS S3** for storing images and chat history.
- Enabled real-time responses by processing both image and text inputs using integrated multimodal capabilities.

Education

M S Ramaiah University of Applied Sciences   **Aug. 2018 – Jun 2022**
Bachelor of Technology in Computer Science and Engineering   *Bengaluru, Karnataka*

Certifications and Publications

- AWS Certified AI Practitioner   Click here to view

- Classification of Ventricular Arrhythmia Using Machine Learning - I4C 2022 by IEEE   Click here to view

- Deepfake detection in digital media forensics - ICIEA-2022 by Science Direct   Click here to view