Venkata Sai Vikas Katuru

Charlotte, NC | vikkyus2772000@gmail.com| (716) 275-7777 | [GitHub](#) | [Linkedin](#)| U.S. work Authorization: (F1-OPT)

SKILLS

Languages: Python, Java, JavaScript

AI/ML: LLMs, NLP, Generative AI, RAG, Transformers, Computer Vision, Agents, RLHF.

Frameworks: PyTorch, TensorFlow, Scikit-learn, Hugging Face, Keras, LangChain, LangGraph, Crew AI,

Databases: MySQL, PostgreSQL, MongoDB, Pinecone

Cloud & DevOps: AWS S3, EC2, SageMaker, Lambda, Azure AI, CI/CD, Git, GitHub Actions, Docker.

Tools/Protocols/methodologies: Flask, FastAPI, REST, React, JWT, Postman, Agile, SCRUM

EXPERIENCE

Data Analyst - Junior| Time series analysis, ARIMA, EDA   Mar 2025

Community Dreams Foundation   United States

- Led exploratory data analysis on 2.5+ years of shipment scheduling data to identify trends, seasonality, and inefficiencies in raw material usage using time series decomposition and correlation analysis.
- Created interactive dashboard to visualize shipment patterns, contributing to a 10% improvement in decision making.
- Collaborated on ARIMA based forecasting to reduce raw material waste by 15–20% through predictive scheduling.

Machine Learning Intern | Crew AI, RAGs, Pinecone, MongoDB, LangSmith   Aug 2024 – Dec 2024

Appetit   United States

- Developed a conversational AI chatbot to enable non-technical stakeholders for querying a MongoDB database using natural language.
- Set up a multi agent system using Crew AI, automating query generation and improving data retrieval speed.
- Optimized product name matching using Pinecone vector embeddings and RAG, improving information retrieval and cutting response latency by 50%.
- Automated schema analysis, reducing query formulation errors by 25%, improving data quality and decision making.
- Monitored agent performance with LangSmith, achieving a 40% improvement in benchmark scores post-optimization.
- Introduced a feedback loop using LLM Judge with scoring and discrepancy checks to iteratively refine agent responses.

PROJECTS

Full-Stack Coding Practice Platform| FastAPI, React +Vite, JWT, MongoDB   Apr 2025

- Created a full-stack coding platform supporting 150+ problems with real-time test case validation.
- Facilitated the scraping and structuring of 150+ problems and 500+ test cases in MongoDB with relevant metadata.
- Built 15+ backend APIs using FastAPI with JWT-secured authentication, and developed a React + Vite frontend featuring a multi-language code editor (Python, JavaScript, Java, Go, C++).

Evaluating Transformers for Clinical Trial Reasoning | Flan T5, Zero-Shot, Fine tuning   Feb 2024 – May 2024

- Collaborated to enhance clinical trial statement assessment by fine-tuning Flan-T5 with 10 instruction templates, producing the most reliable model and earning team recognition for experimental contributions.
- Achieved a top F1-score of 0.78 with the best model, contributing insights to clinical trial assessment research.

Obesity Risk Prediction using Deep Learning | Flask, AWS, Neural Network.   Feb 2024 – May 2024

- Designed and deployed a neural network-based obesity risk predictor as a real-time Flask service on AWS EC2 using medical and demographic inputs.
- Optimized the Neural Network architecture, achieving 92% accuracy and outperforming SVM and Logistic Regression.

Image Inpainting & Super-Resolution| Context Encoder, U-net   Feb 2024 – May 2024

- Devised a U-Net based image inpainting model to fill missing parts of CIFAR-10 images, achieving PSNR of 29.8 and SSIM of 0.91 on test data.
- Compared U-Net and Context Encoder, selecting U-Net for superior reconstruction on low resolution images (32 X 32).

EDUCATION

State University of New York at Buffalo   Aug 2023 – Jan 2025

School of Engineering and Applied Sciences: Masters in Artificial Intelligence   Buffalo, NY

G. Pulla Reddy Engineering College   Jul 2018 – Jun 2022

Bachelor of Technology in Mechanical Engineering   India