KANTIPUDI KESAVA SAI VEERENDRA

RAJAHMUNDRY | kantipudikesavasaiveerendra@gmail.com | 8790418895 | kesavakantipudi.vercel.app |
linkedin.com/in/kesavakantipudi | github.com/kesavakantipudi

SUMMARY

AI Engineer specializing in building real-world intelligence systems using LLMs and machine learning. Experience in developing Retrieval-Augmented Generation pipelines, scalable AI APIs, real-time analytics platforms, and Ai safety middleware. Strong focus on making AI reliable, and production-ready rather than only model training.

Technical Skills

Programming: Python, SQL
Machine Learning: Classification, Clustering, Feature Engineering, Model Evaluation
Deep Learning & LLM: Transformers, LoRA Fine-Tuning, Prompt Engineering
GenAI Systems: RAG Pipelines, Vector Databases, Embeddings, Guardrails
Backend & Deployment: FastAPI, REST APIs, Async Processing, Docker
Data & Tools: NumPy, Pandas, Scikit-learn, HuggingFace, Redis, PostgreSQL
Cloud: AWS(EC2, S3)

Education

B.Tech - AI & ML, Aditya College of Engineering & Technology — 2023 - 2027 | GPA: 8.5/10
Intermediate (MPC), Aditya Junior College — 2021 - 2023 | 89.1%

Experience

Data Specialist Intern - Technical Hub — *May 2025 - June 2025*

- Built data processing workflows and analytics pipelines
- Automated reporting systems reducing manual effort
- Worked with real datasets to generate actionable insights

Projects

RAG-Based Document Question Answering System GITHUB

- Built a semantic search pipeline using embeddings and vector database retrieval
- Generated grounded responses to eliminate LLM hallucinations
- Implemented document chunking, retrieval ranking, and citation generation
- Designed modular architecture simulating enterprise knowledge assistants

Scalable AI Inference Service GITHUB

- Designed asynchronous ML inference API using FastAPI and worker queues
- Implemented caching and rate limiting to reduce compute cost
- Separated request handling from model execution for scalability
- Production -style architecture supporting concurrent users

Prompt Injection Defence system for LLM Applications GITHUB

- Built middleware detecting prompt injection and jailbreak attempts
- Sanitized inputs and outputs to prevent sensitive data leakage
- Designed reusable wrapper for safe LLM deployment

Certifications

AWS Certified AI Practitioner (Certificate)

NPTEL: Machine Learning (Certificate)

Oracle: Java Certified Foundations Associate (Certificate)

Oracle: Certified Foundations Associate — Database (Certificate)

SnowProAssociate: Platform (Certificate)

Certiport: IT Specialist - HTML and CSS (Certificate)

Achievements

Solved **800+** coding problems accross **LeetCode, CodeChef, GeeksforGeeks and HackerRank**

Strong problem-solving and algorithm thinking foundation