Sai Supraja Ravi Kumar

Houston, Texas | +1-551-310-1086 | saisuprajaravikumar.2025@gmail.com | Linkedin | GitHub | Portfolio

Summary

AI/ML Engineer with **4+ years of experience** building and deploying **machine learning, deep learning, and Generative AI systems** across **healthcare, finance, and enterprise analytics**. Proven ability to design **end-to-end ML and LLM-powered platforms**-from data pipelines and model development to **cloud-native deployment and full-stack applications**. Strong background in **predictive modeling, NLP, RAG architectures, and scalable MLOps**, delivering solutions that improved decision accuracy by **20%+**, reduced research time by **70%**, and automated critical workflows. Passionate about building **reliable, explainable, and production-grade AI systems**.

Education

University of Houston **Aug 2023 - May 2025**
Master of Science, Data Science
- **GPA:** 3.98

Anna University **2018 - 2022**
Bachelor of Engineering, Electronics & Communication
- **GPA:** 8.68

Professional Experience

Optum **May 2024 - Present**
AI/ML Engineer – Generative AI & Applied Machine Learning
- Designed and deployed **LLM-powered RAG applications** using OpenAI, LangChain, FAISS, and AWS to enable contextual search and question answering over structured and unstructured enterprise datasets, **reducing analyst research time by 65–70%**.
- Built and productionized **full-stack GenAI applications** using **React, FastAPI, and containerized backends**, delivering interactive AI assistants with **~90% factual accuracy**.
- Developed **machine learning models** (Random Forest, Gradient Boosting, Neural Networks) on large-scale healthcare and financial datasets to support **risk prediction and decision support**, improving model precision and downstream decisions by **20%+**.
- Architected scalable **data and inference pipelines** using **Databricks, Spark, FAISS, Docker, and Kubernetes**, supporting **9M+ records** with low-latency retrieval.
- Deployed end-to-end ML and GenAI workflows on **AWS (Lambda, SageMaker, API Gateway, CloudWatch)** and Azure, reducing deployment latency by **40%**.
- Implemented **model monitoring, logging, and CI/CD pipelines**, collaborating with cross-functional teams through code reviews and design discussions to ensure production reliability and scalability.
- Applied **model explainability techniques (SHAP)** and governance best practices to support compliance, transparency, and trust in AI-driven decisions.

University of Houston **Dec 2023 - May 2025**
Generative AI Researcher – Biomedical & Workflow Automation *Houston*
- Built a **biomedical literature RAG system** using Hugging Face Transformers, FAISS, and Azure OpenAI, enabling fast semantic retrieval over PubMed-scale corpora and **cutting domain query time by 70%**.
- Designed **agentic AI workflows** using LangChain and LangGraph to orchestrate multi-step reasoning, retrieval, and summarization pipelines.
- Containerized and deployed **multi-turn LLM assistants** using Docker and Azure Functions, achieving **82% answer accuracy** on biomedical QA benchmarks.
- Developed **LLM evaluation and experimentation pipelines** using Streamlit, MLflow, and HELM-style metrics, improving model performance tracking and iteration speed by **28%**.
- Applied **prompt engineering and schema-guided generation** to automate scientific documentation and reporting, saving **400+ engineering hours**.
- Implemented backend services using **Python and FastAPI** for real-time QA and document intelligence systems.

Capgemini **Apr 2021 - Jun 2023**
Data Analyst – ML & Full-Stack Analytics Systems *Chennai, India*
- Designed and deployed **time-series forecasting models (ARIMA, LSTM)** for insurance claims and financial transactions, achieving **92% forecast accuracy**.
- Built **end-to-end ETL pipelines** using Python, SQL, EventHub, and Airflow to process **3M+ daily records**, enabling reliable downstream analytics and ML workloads.
- Developed **classification and NLP pipelines** using Scikit-learn and spaCy to triage unstructured claims and financial text, **reducing manual review effort by 50%**.
- Created **RESTful APIs** and deployed ML services using Docker and AWS SageMaker to support real-time anomaly detection, improving payment risk alerts by **33%**.

- Designed **Power BI dashboards and embedded analytics** to surface KPIs across **50+ payment and claims systems**, accelerating data-driven decision-making.

Projects

GenAI-based Text Summarization System - Abstractive NLP with LLMS **Oct 2024 - Nov 2024**

- Designed an advanced text summarization pipeline using transformer-based models (BERT, Pegasus) on CNN/DailyMail dataset, achieving a 20% ROUGE score improvement.
- Incorporated self-attention mechanisms and encoder-decoder architecture to refine summarization quality for real-time customer interactions.

ARPU-Driven Strategic Merger Analysis Between Telecom Providers **Jan 2025 - Apr 2025**

- Created interactive dashboards to simulate the financial impact of a merger, showing a projected 18% uplift in ARPU and customer base expansion by 1.3 million.

AI-Driven Food Delivery Chatbot Using Dialogflow & FastAPI **Jun 2024 - Aug 2024**

- Built a chatbot with NER-enhanced real-time interactions, powered by FastAPI achieving 100% accuracy in order management.

Electric Vehicle Market Sales Analysis **Mar 2024 - May 2024**

- Led a End to end ETL pipeline to analysis and forecast demands of EV sales and provide informed insights building a power BI dashboard enhancing decision making by 60%.

Professional Skills

- **Domain/Functional Areas**: Data Analysis & Business Intelligence, Big Data & Data Engineering, NLP & Generative AI, AI/ML Engineering, Data Science
- **Programming Skills**: Python, SQL, R, Java, C++, Matlab, Julia, Javascript, PHP, SAS, PostgreSQL, React
- **Data Analysis & Visualization**: Power BI, Tableau, Excel, Web Scraping, Pandas, Scikit-learn, Data Mining
- **Machine Learning & AI**: Pytorch, TensorFlow, Keras, NLTK, JAX, Hugging Face, Deep Learning, langChain, LLMs, RAG, Prompt Engineering, Vector databases, Azure OpenAI
- **Cloud, Big Data & DevOps**: AWS, Azure AI, GCP, Apache Spark, Hadoop, Databricks, MLflow, Snowflake, Docker, Kubernetes, Cloud Platforms, AWS Sagemaker, AWS Lambda, DevOps
- **Technical & Automation Tools**: Git, Jupyter Notebook, Google Colab, PEGA, A/B Testing, WhyLabs, Langtest, REST APIs
- **Professional Attributes**: Quick Learner, Business Acumen, Proactive Thinking, Software Engineering, Willingness To Learn
- **Quantitative & Analytical Skills**: Mathematics, Time Series Analysis, Quantitative Finance, A/B testing

Certifications

- **Oracle Certified Generative AI: Oracle**
- **Oracle Cloud Infrastructure AI Foundations Associate: Oracle**
- **Introduction to Deep Learning: Microsoft**
- **Data Analytics : Deloitte**