SAI KUMAR REDDY

Phone: +1 (972) 710-9106 | Email: reddysaikumar614@gmail.com

PROFESSIONAL SUMMARY

AI/ML Engineer with 3 years of experience designing and deploying intelligent machine learning and Generative AI solutions across customer analytics and enterprise support automation. Experienced in building **LLM-powered applications**, **Retrieval-Augmented Generation (RAG) pipelines**, semantic search systems, and Transformer-based NLP models using **LangChain, Hugging Face, PyTorch, and OpenAI APIs**. Skilled in developing scalable AI platforms with **Pinecone, FAISS, ChromaDB, Docker, Kubernetes, AWS SageMaker, and MLflow**, enabling automated knowledge retrieval, production model deployment, and continuous monitoring. Strong foundation in classical machine learning, MLOps, and cloud-native AI systems with a proven ability to translate business challenges into production-ready AI solutions within Agile environments.

TECHNICAL SKILLS

- **Programming Languages:** Python, SQL, Java
- **Generative AI & LLMs:** LangChain, Hugging Face Transformers, Retrieval-Augmented Generation (RAG), Prompt Engineering, OpenAI API, LlamaIndex, LangGraph, AI Agents
- **Machine Learning & Deep Learning:** PyTorch, TensorFlow, Scikit-learn, Keras, CNNs, RNNs, Transformers, Transfer Learning
- **Natural Language Processing (NLP):** BERT, spaCy, Tokenization, Text Classification, Sentiment Analysis, Embedding Models, Semantic Search
- **Vector Databases:** Pinecone, FAISS, ChromaDB
- **MLOps & Model Deployment:** Docker, Kubernetes, MLflow, FastAPI, REST APIs, GitHub Actions, CI/CD, Model Monitoring
- **Cloud Platforms:** AWS SageMaker, Microsoft Azure, Google Cloud Platform (GCP)
- **Data Engineering & Processing:** Pandas, NumPy, ETL Pipelines, Feature Engineering, Data Preprocessing
- **Computer Vision:** OpenCV, Image Classification
- **Version Control & Development Tools:** Git, GitHub, Jupyter Notebook, Linux
- **Model Evaluation:** Cross-Validation, Hyperparameter Tuning, Confusion Matrix, A/B Testing, Performance Optimization
- **Methodologies:** Agile/Scrum, Technical Documentation, Problem Solving, Team Collaboration

PROFESSIONAL EXPERIENCE

Teledyne Technologies Inc | AI/ML Engineer | Jul 2025 – Present

- Designed and developed an **Intelligent Technical Knowledge and Support Automation Platform** to automate technical support by leveraging **Retrieval-Augmented Generation (RAG)** and enterprise knowledge sources.
- Built an end-to-end **RAG pipeline** using **LangChain**, **OpenAI API**, **FAISS**, and **Pinecone** to retrieve, rank, and generate context-aware responses from technical manuals, SOPs, and support documentation, reducing support ticket volume by **35%**.
- Fine-tuned **Hugging Face Transformer** models, including **BERT**, for intent classification and technical query understanding, improving response accuracy by **18%**.
- Enhanced semantic search capabilities by generating and managing vector embeddings in **Pinecone** and **ChromaDB**, improving search relevance by **30%** and reducing retrieval latency.
- Built reusable prompt engineering frameworks and implemented **agentic workflows** using **LangGraph** and **LlamaIndex** to orchestrate retrieval, reasoning, and tool execution for complex technical support requests.
- Containerized AI services using **Docker** and deployed scalable inference workloads on **Kubernetes** and **AWS SageMaker**, reducing deployment time by **40%** and improving production scalability.
- Automated model training, deployment, and version management using **GitHub Actions**, **CI/CD**, and **MLflow**, enabling continuous model monitoring, experiment tracking, and proactive retraining based on data drift.
- Collaborated with software engineers, product teams, and domain experts to integrate AI capabilities into enterprise applications while mentoring junior engineers through code reviews and technical guidance.

Colt Technology Services | Junior Machine Learning Engineer | May 2021 – Jun 2023

- Contributed to the development of a **Customer Retention Intelligence and Churn Prediction Platform** to identify at-risk customers and support data-driven retention strategies.
- Collected, cleaned, and transformed **250K+ customer records** using **Pandas** and **NumPy**, improving data quality and reducing pre-processing time by **30%**.
- Performed feature engineering and built supervised machine learning models using **Scikit-learn** and **TensorFlow**, improving churn prediction accuracy by **15%**.
- Improved ETL pipelines to integrate customer usage, billing, and support data from multiple enterprise systems into machine learning workflows.
- Evaluated and optimized classification models using **cross-validation**, confusion matrix, precision, recall, and F1-score to improve prediction reliability.

- Assisted in deploying trained models as **FastAPI REST APIs** on **AWS**, reducing inference response time by **20%** and enabling real-time churn prediction.
- Collaborated with data engineering and business teams to monitor model performance, maintain Linux-based ML environments, and continuously improve feature engineering pipelines within Agile development cycles.

EDUCATION

- **Master of Science in Advanced Data Analytics | August 2023 – May 2025**
  University of North Texas Denton, TX
- **Bachelor of Science in Computer Science | June 2018 – May 2021**
  Osmania University Hyderabad, India

CERTIFICATION

- **AWS Certified Solutions Architect - Associate**