SREE CHARAN REDDY KAILASAM

AI/ML Engineer

sreecharanreddy90@gmail.com | +1 (443) 855-2775 | LinkedIn

SUMMARY

AI/ML Engineer with **3+ years** of experience designing, optimizing, and deploying production-grade machine learning, deep learning, and generative AI solutions within large-scale enterprise frameworks. Expert in engineering high-performance retrieval-augmented generation (RAG) pipelines, custom multi-agent execution graphs, and automated MLOps pipelines across AWS, Azure, and GCP. Proven track record of optimizing real-time inference clusters, tuning distributed storage architectures, and orchestrating containerized microservices to maximize application scalability and maintain high-availability system SLAs.

SKILLS

- **Programming & Core Frameworks:** Python, SQL, R, Scikit-learn, XGBoost, Pandas, NumPy
- **Deep Learning & Generative AI:** PyTorch, TensorFlow, Keras, Hugging Face, OpenAI API, LangChain, LlamaIndex
- **LLM Engineering & Vector Search:** Prompt Engineering, RAG Architecture, Fine-tuning, Pinecone, FAISS, Weaviate
- **Cloud Platforms & Infrastructure:** AWS (SageMaker, Lambda, S3, EC2), Azure ML, GCP Vertex AI
- **Distributed Computing & Big Data:** Apache Spark, PySpark, Apache Kafka, Hadoop, Databricks
- **MLOps, DevOps & Orchestration:** MLflow, Kubeflow, Docker, Kubernetes, Airflow, Jenkins, GitHub Actions
- **Data Warehousing & Lakehouses:** ETL Pipelines, Snowflake, Amazon Redshift, Delta Lake
- **Relational & NoSQL Databases:** PostgreSQL, MySQL, MongoDB, Cassandra
- **Core APIs & Observability Stack:** REST APIs, FastAPI, Flask, Prometheus, Grafana
- **Data Visualization & Analytics:** Power BI, Tableau, Matplotlib, Seaborn

EXPERIENCE

AI/ML Engineer| OpenAI, USA **Jun 2025 – Present**

- **Business Impact:** Automated contextual enterprise search across major business lines, cutting manual data discovery times and lowering production support ticket backlogs.
- Designed and deployed LLM-based enterprise applications using GPT APIs and open-source models (LLaMA, Mistral), improving contextual response accuracy across specialized domains.
- Built Retrieval-Augmented Generation (RAG) pipelines using LangChain and Pinecone, **reducing model hallucination rates by 35%** on core verification benchmarks.
- Developed scalable AI inference services using FastAPI and Kubernetes to handle **over 1M+ daily user requests** across high-concurrency clusters.
- Optimized prompt engineering strategies and few-shot templates, leading to steady improvements in downstream model evaluation relevance scores.
- Implemented a distributed vector search architecture using FAISS and Weaviate to execute low-latency semantic search over 50M+ unstructured documents.
- **Reduced real-time model inference latency thresholds by 60%** using INT8 model quantization and customized distributed caching strategies.
- Built automated MLOps production pipelines using MLflow and AWS SageMaker, streamlining end-to-end cloud model deployment workflows.
- Collaborated with cross-functional software engineering teams to smoothly integrate production AI SaaS interfaces with a verified 99.9% application uptime.
- **Technologies:** OpenAI APIs, GPT-4, LLaMA, Mistral, LangChain, Pinecone, FAISS, Weaviate, FastAPI, Kubernetes, MLflow, AWS SageMaker

Data Scientist| Nvidia, India **Jun 2021 – Dec 2023**

- **Business Impact:** Upgraded predictive telemetry performance and data extraction workflows, reducing operational hardware compute overhead footprints.
- Developed predictive models using XGBoost, Random Forest, and deep learning methods in TensorFlow, **improving diagnostic baseline accuracy parameters by 22%**.
- Built end-to-end big data processing pipelines using Apache Spark, PySpark, and RAPIDS, **transforming over 10TB+ of distributed telemetry data daily**.
- Designed real-time monitoring dashboards using Power BI and Tableau, streamlining internal stakeholder metric reporting turnaround cycles.
- Implemented anomaly and fraud detection systems using ensemble techniques, successfully **decreasing system infrastructure false positive alerts by 28%**.
- Managed automated cloud data workflows across AWS Glue, S3, and Hadoop distributed filesystems to optimize backend cluster processing scalability.
- Developed product recommendation systems using neural collaborative filtering architectures built on Keras, boosting user digital engagement and platform retention.
- Automated training loops using Kubeflow, Jenkins, and GitHub Actions to minimize code regression bottlenecks across delivery cycles.

- Executed deep feature engineering, selection matrices, and PCA dimensionality reduction in R and Python to stabilize volatile predictive inputs.
- **Technologies:** TensorFlow, Keras, Apache Spark, PySpark, RAPIDS, AWS (S3, Glue), Kubeflow, Jenkins, GitHub Actions, R, Power BI, Tableau, SQL

PROJECTS

Real-Time Financial Fraud Detection Platform

- Designed a streaming anomaly detection platform using Apache Spark, Kafka, and SQL to ingest and parse high-frequency transactional data records.
- Applied ensemble configurations via XGBoost and Scikit-learn, achieving a verified 30% improvement in precise fraud flag captures.
- Processed over 10M+ training rows while scaling deployment architectures on AWS SageMaker to lower infrastructure evaluation delays by 22%.
- Engineered automated data ingestion tasks using Airflow to pipe cleared transaction vectors directly into a centralized Redshift data warehouse.
- **Technologies:** Python, Apache Spark, Apache Kafka, XGBoost, Scikit-learn, AWS SageMaker, Airflow, Amazon Redshift, SQL

Context-Aware Customer Support Engine

- Developed an AI chatbot utilizing Hugging Face Transformers and NLP tokenizers to automate high-concurrency customer query analysis.
- Fine-tuned a transformer BERT model for intent classification, increasing baseline classification accuracy by 25% and reducing resolution times by 30%.
- Deployed containerized microservices via FastAPI and Docker to AWS EC2 instances, supporting real-time concurrent chat connections.
- Integrated text sentiment classification routines to dynamically adjust conversational prompt contexts based on real-time user emotion scores.
- **Technologies:** Python, Hugging Face Transformers, BERT, NLP, FastAPI, Docker, AWS EC2, Prompt Engineering

Multi-Cloud Enterprise Knowledge Mesh & Observability Layer

- Built a decentralized multi-index RAG system using LlamaIndex to orchestrate data transformation paths across Snowflake and Delta Lake lakehouses.
- Trained deep neural indexing layers across Azure ML and GCP Vertex AI frameworks, expanding multi-cloud search accuracy parameters by 35%.
- Constructed real-time microservices via Flask to extract payload records while maintaining zero data-loss over Cassandra and MongoDB instances.
- Configured infrastructure monitoring architectures using Prometheus and Grafana, reducing system pipeline debugging turnaround times by 40%.
- **Technologies:** LlamaIndex, Azure ML, GCP Vertex AI, Snowflake, Delta Lake, Flask, Prometheus, Grafana, Cassandra, MongoDB

EDUCATION

Master of Science in Data Science | University of Maryland, Baltimore County (UMBC), Baltimore, USA **Dec 2025**

CERTIFICATION

- **AWS Certified Machine Learning – Specialty** | Amazon Web Services
- **Microsoft Certified: Azure AI Engineer Associate** | Microsoft
- **Google Cloud Certified Professional Machine Learning Engineer** | Google