Sai Krishna Sriram

San Francisco, CA | Open to Relocation
Phone: (720) 233-6444 | **Email**: saikrishna.sriram3@gmail.com
LinkedIn: https://www.linkedin.com/in/saikrishnasriram |**GitHub**: https://github.com/Krish3na
Portfolio: https://saikrishnasriram.netlify.app

Summary

Generative-AI and ML Engineer who ships production LLM systems with measurable impact: conversational RAG agents, multi-agent workflows, and real-time analytics on AWS, Snowflake, and Databricks. Built chatbots with grounded citations, evaluation frameworks, and serverless OCR/data apps, always with safety rails (validation, timeouts, rate limits), CI/CD, and observability to turn prototypes into reliable products.
Strengths: GenAI Systems Design, Production-grade LLM delivery, RAG evaluation and tuning, MLOps and observability, Streaming and Data Engineering, Clear, impact-focused communication.

Technical Skills

GenAI & ML: Python, R, PyTorch, TensorFlow, OpenAI API, Llama API, embeddings, prompt engineering, RAG, LangChain, LangGraph, CrewAI, LlamaIndex, vector DBs (ChromaDB, FAISS), evaluation (precision@k, NDCG), hallucination reduction, NLP, scikit-learn
MLOps and Backend: FastAPI, Flask, gRPC, MLflow, Docker, Kubernetes, GitHub Actions, Jenkins, blue-green and rollback, monitoring and alerts (Prometheus, Grafana), Git, CI/CD, DevOps
Data and Compute: Python, SQL, PySpark and Databricks, Kafka, Airflow, Hive, ETL and ELT, streaming
Cloud: AWS (Lambda, Glue, Step Functions, S3, Kinesis, RDS, EC2, ECS, EMR, SageMaker, Bedrock, Lex, Athena, Redshift), Azure, GCP (Vertex AI, Cloud Run, GKE, AppSheet)
Warehousing and BI: Snowflake, Power BI, Tableau, QlikView, Quicksight

Work Experience

CLD-9 **Temecula, California**
AI/ML Engineer **Jul 2025 - Present**

- Productionized an LLM supplement recommender as a FastAPI service on AWS with MLflow and CI/CD; added input validation, rate limits, timeouts, and safe rollback to harden reliability.
- Built a serverless OCR pipeline: S3 events to EventBridge to Step Functions to Lambda (Pixtral with PyMuPDF), delivering sub-minute processing for multi-page blood reports.
- Implemented an insights layer to normalize lab ranges, flag out-of-range values, and map findings to precautions and supplements; retries and failure thresholds reduce reprocessing and error loops.

AI/ML Intern **Jan 2025 - Jun 2025**

- Prototyped a RAG-based supplement recommendation engine (OpenAI and LangChain over vetted medical sources) with interactive follow-ups and grounded citations; 35% relevance lift and 20% fewer hallucinations.
- Built retrieval/ranking with OpenAI embeddings, contraindication and business-rule filters, and prompt templates.
- Built an evaluation loop (precision@k, NDCG, factuality checks, user feedback) that cut manual review by 60%.

ASANTe **San Francisco, California**
AI Engineer Intern (Capstone) **Jan 2025 - May 2025**

- Shipped Flask and gRPC NLP services on AWS Lambda; integrated text-embedding-3-large and improved recommendation accuracy by 18%.
- Built hybrid recommenders (collaborative, content-based, sequential/BST) across retail and nonprofit use cases.
- Orchestrated Airflow pipelines for ingest to embeddings to training to evaluation and drift monitoring, with Tableau KPI reports.

AbsoluteLabs. **Hyderabad, India**
Associate Consultant (ETL Developer) **Sep 2022 - Jul 2023**

- Engineered Databricks PySpark and Hive pipelines processing 5M+ records/day; optimized joins and aggregations to reduce ETL runtime by 20% and meet SLAs.

- Created feature-ready datasets in Snowflake; advanced SQL (CTEs and window functions) surfaced demand forecasts and stock-out risk; shipped Power BI store-level KPI dashboards.
- Added Prometheus and Grafana monitoring and schema-change alerts; built data-quality checks for volume, freshness, and validity.

Tata Consultancy Services **Hyderabad, India**
Assistant System Engineer (Data Engineer – Client: Qualcomm) **Aug 2021 - Aug 2022**

- On-site at Qualcomm, built an ML analytics pipeline (Kafka and Scala to Elasticsearch, archive on S3 Glacier) with autoencoder-based anomaly detection; QlikView dashboards reduced failure-triage time by ~30%.
- Developed Python TCP/UDP diagnostics (iPerf) to quantify latency and throughput; containerized services in Docker and Kubernetes with Jenkins CI/CD for global releases.

Career Launcher - Aspiration AI **New Delhi, India**
Machine Learning Intern (Remote) **Apr 2020 - Jun 2020**

- Built a pandas equity-data pipeline (OHLCV ingest, VWAP, rolling volatility, SMA and Bollinger) and trained a Random Forest classifier with out-of-sample evaluation; ran cumulative-return backtests and portfolio optimization (efficient frontier) and validated diversification via K-means.

Projects

AWS Bedrock Conversational RAG Agent | https://github.com/Krish3na/aws/bedrock/conversational-rag-agent

- Built a production-style conversational agent using Amazon Bedrock Knowledge Bases and Agents; ingested documents from S3, generated and stored embeddings in Aurora PostgreSQL via Amazon Titan Embeddings, and orchestrated secure question-answering with Amazon Nova LLM, Secrets Manager, and VPC-integrated infrastructure.

Kaggle Multi-Agent System (LangGraph) | https://github.com/Krish3na/kaggle-multi-agent-system

- Autonomous Kaggle system using LangGraph state graphs and LLM agents (GPT-4/Claude) for intelligent decision-making; integrated LangChain tools for API operations, data processing, and ML training (LightGBM/XGBoost/CatBoost) with MLflow tracking.

GenAI Chatbot (LangChain + GPT-4) | https://github.com/Krish3na/genai-chatbot

- RAG over domain docs using OpenAI, LangChain, and ChromaDB, exposed via FastAPI (Docker) and monitored with MLflow; ~30% higher factual accuracy, ~25% lower cost, and ~60% faster scaling.

Real-Time Expense Tracker | https://github.com/Krish3na/Real-TimeExpenseTracker

- Kinesis to Lambda to RDS (PostgreSQL), receipts to S3, Flask API, React dashboard on EC2; processes transactions every 2 seconds with MCC-based categorization and fraud-risk scoring; CloudWatch monitoring and IAM controls.

Churn Prediction and Retention Reporting | https://github.com/Krish3na/churn-prediction

- SQL and Python feature pipeline and Random Forest (94.05% accuracy) with Streamlit and Power BI dashboards to target the top 9.1% at-risk customers.

Education

M.S., Data Science - University of Colorado Boulder

B.Tech, Electronics & Communication Engineering – GRIET (Hyderabad)

Certifications

Generative AI with Large Language Models | Coursera - Built and deployed end-to-end dialogue summarization system (prompting, LoRA PEFT, RLHF with PPO, **SageMaker** deployment).

AWS Certified Data Engineer - Associate | Credly

Data Analytics & Machine Learning Certificate | ElevateMe