Sai Kiran

Texas, United States saikiran.sorce@gmail.com 4692902262

SUMMARY

Computer Science graduate and **AWS Certified Solutions Architect – Associate** with expertise in software development, cloud computing, and ML/Gen AI. Proficient in Python and Java, with hands-on experience building and deploying **scalable full-stack and ML applications** on AWS and Linux. Experienced **Graduate Teaching Assistant**, mentoring 50+ students on programming projects. Skilled in **ML pipelines, LLM/Gen AI workflows, and cloud deployment**, passionate about delivering high-quality, impactful engineering solutions.

PROFESSIONAL EXPERIENCE

ML Engineer

Texas • Deltasoft Solutions • March 2025 - Present

- Contributed to developing an **enterprise GenAI chatbot** using **OpenAI GPT APIs**, **LangChain**, and **FAISS**, helping employees query internal documents and reduce search time by **~40%**.
- Assisted in building **RAG pipelines** integrated with **Azure Cognitive Search** to deliver context-aware responses from **financial and compliance reports**.
- Supported **fine-tuning of LLaMA-2 (7B)** and **GPT-J** models on **banking datasets**, improving domain-specific accuracy by **~30%**.
- Helped implement **intent detection, context tracking, and fallback strategies**, increasing multi-turn conversation reliability by **~25%**.
- Co-developed **Streamlit dashboards** for real-time monitoring of **latency, token usage, and response quality metrics**.
- Collaborated with senior engineers to integrate **feedback loops, validation checks, and error logging**, ensuring consistent and auditable chatbot outputs.

Data Analyst

Hyderabad, India • Osair Technologies • January 2022 - December 2022

- Cleaned and processed **50k+ transactional and behavioral records** using Python (Pandas, NumPy), improving data readiness for downstream modeling by ~30%.
- Assisted in building and evaluating **baseline ML models** (linear/logistic regression, decision trees, clustering), boosting prediction accuracy by ~15–20% compared to heuristics.
- Designed and maintained **data preprocessing pipelines** for 50k+ daily records, handling missing data, scaling, and feature engineering.
- Implemented **MLflow for experiment tracking** and dataset/model versioning.
- Built **Streamlit dashboards** to visualize model metrics, enabling non-technical teams to track model performance.
- Deployed ML models as REST APIs via **FastAPI**, improving accessibility for internal applications.

EXPERIENCE

Graduate Teaching Assistant

George Mason University **January 2024 - January 2025, Fairfax, VA**

- Supported instruction for a programming course, helping over 50 students strengthen their understanding of object-oriented programming and foundational CS concepts.
- Led discussions and assisted during lab sessions, breaking down complex topics into simpler terms, which contributed to a noticeable improvement in student performance.
- Reviewed more than 100 coding assignments, offering clear, constructive feedback to encourage better coding practices and logical problem-solving.
- Developed strong communication and mentoring skills, reinforcing a collaborative learning environment and sharpening technical leadership qualities.

CERTIFICATIONS

AWS Certified Solutions Architect - Associate | 2025

https://www.credly.com/badges/d7730cb7-d30e-4d7d-9ca7-ac7dc25b148d/public_url

AWS Certified Cloud Practitioner | 2025

https://www.credly.com/badges/10597c59-b4d4-41d7-8a75-1c6d1a7f0e20/public_url

SKILLS

Programming Languages: Python (PyTorch, NumPy, Pandas, Scikit-learn) , Java, C++, C, JavaScript
Generative AI: OpenAI APIs, LangChain, Retrieval Augmented Generation (RAG), Vector Databases (FAISS, Pinecone)
Machine Learning & AI: Supervised and Unsupervised Learning, Neural Networks, Model Evaluation, Data Visualization
Web & Frontend: HTML, CSS, React, PHP
Cloud & DevOps: AWS (EC2, S3, RDS, DynamoDB, Lambda, SNS), Docker
Databases: MySQL, Amazon RDS, DynamoDB
Operating Systems: Linux, Windows
Networking & Security: TCP/IP, DNS, Routing, Firewalls, API security
Tools & IDEs: Git, Streamlit, FastAPI, Android Studio, NetBeans

EDUCATION

Master of Science in Computer Science
George Mason University • Fairfax, VA • 2024 • 3.90

Bachelor of Technology in Computer Science and Engineering
Sri Indu College of Engineering and Technology • Hyderabad, India • 2022 • 3.70

PROJECTS

Smart Document Q&A System | Python, PyTorch, RAG, LangChain

George Mason University • August 2024 - December 2024

- Built a **document retrieval Q&A system** using Hugging Face Transformers, LangChain, and FAISS, enabling users to query 1,000+ PDF files with context-aware answers in under 2 seconds.
- Fine-tuned a **pre-trained LLM with LoRA** on domain-specific data, improving response accuracy by **27% compared to baseline zero-shot performance**.
- Deployed the system with **FastAPI and Docker**, creating an interactive web app that handled **50+ concurrent queries** with <200ms latency on average.

Online Banking System | Python, JavaScript, CSS, MySQL, AWS

George Mason University • January 2024 - May 2024

- Built a **full-stack banking app** with a login page for customer and employee; supporting account creation, deposits, withdrawals, transfer between accounts, balance checks and account merging.
- Designed a normalized **MySQL schema** with ACID transactions to ensure **100% data consistency** across user accounts.
- **Hosted on AWS**: Backend on **EC2 (Linux)**; static frontend on **S3**; access controlled via **security groups**.

Mobile App and Website for Frequent Flyer Program | Java (Android), HTML/CSS/JavaScript, AWS, RDS

George Mason University • January 2023 - May 2023

- Developed a **Java-based Android app and responsive web interface** for an airline frequent flyer program, serving users to manage accounts, book flights, and redeem points.
- Designed and deployed **scalable backend services on AWS**, using Amazon RDS for secure storage of user and booking data, handling **100+ transactions/day** with high availability.
- Integrated mobile and web frontends via **REST APIs**, ensuring real-time data sync, modular design, and end-to-end full-stack functionality.