Udayan Atreya

San Jose, CA | 669-340-6033 | udayan.atreya@sjsu.edu | LinkedIn - /uatreya | Github | Portfolio

EDUCATION

Master of Science, Computer Science Expected May 2026
San Jose State University, San Jose, US
Coursework: Retrieval Augmented Generation (RAG), Applied Deep Learning, Biometric Security with AI.

Bachelor of Technology, Computer Science & Engineering May 2020
Manipal University, Jaipur, India
Coursework: Data Structure and Algorithms, Relational Database Management Systems, OOPS.

TECHNICAL SKILLS

Frameworks: LangChain, HuggingFace, TensorFlow, PyTorch, Git, LLM, RAG, FastAPI, RAGAS
Cloud Technology: AWS, S3, Lambda, API Gateway, Azure, CI/CD pipelines, MongoDB, DynamoDB, Linux
Languages: Python, Java, C#, SQL, PowerShell, HTML, CSS.

PROJECT EXPERIENCE

Multimodal RAG: Comparative Study & Evaluation Aug 2025 – Present
- Developed multimodal Retrieval-Augmented Generation (RAG) system supporting text, audio, and video, enabling cross-modal knowledge retrieval through **vector embeddings and similarity-based search.**
- **Evaluated system performance** using RAGAS metrics - achieving Faithfulness: 0.81, Context Precision: 0.78, Context Recall: 0.88, and Answer Relevancy: 0.84 - ensuring contextually grounded responses.
- Benchmarked multiple LLMs (Gemini, DeepSeek, Llama, GPT, Ollama), embedding models and similarity strategies, yielding **13% improvement in retrieval precision** over baseline implementations.

Agentic AI for Infrastructure Observability & Control Jun 2025 – Jul 2025
- Built an **end-to-end observability** and automation platform for data center hardware using **agentic LLMs** and telemetry analytics.
- Engineered a modular architecture integrating **FastAPI, MongoDB, Prometheus, Grafana, LangChain, and Gemini 1.5** to monitor server telemetry and trigger self-healing actions using natural language.
- Designed and deployed a React-based dashboard providing **real-time monitoring** of 20+ hardware metrics (CPU, GPU, RAM, network I/O, disk usage) [GitHub Link].

Human Activity Recognition with Deep Learning Jan 2025 – May 2025
- Benchmarked **Decision Tree, K-NN, Naive Bayes, SVM, Random Forest, CNN, and BiLSTM** on multimodal time series dataset classifying 6 activities via 5-fold cross-validation.
- Implemented **feature engineering** to extract 20+ time and frequency domain features, **improving model accuracy by 15%** over raw signal baselines.
- Applied **LOSO (Leave-One-Subject-Out) transfer learning** using HAR-CNN architecture, boosting generalization performance **accuracy by 10%** [GitHub Link].

WORK EXPERIENCE

Software Engineer, Accenture, Bengaluru, India Dec 2021 – Jul 2024
Associate Software Engineer, Accenture, Bengaluru, India Nov 2020 – Dec 2021
- Engineered custom CI/CD task **orchestrating 5+ microservices** for remediating code vulnerabilities; harnessing AWS serverless architecture and .NET; **safeguarding 400+ applications** over 3+ orgs.
- **Decreased manual ops labour up to 60%** by enabling chatbot integrations for 100+ services; developed modules, orchestration flows and led cross-functional collaboration to onboard 20+ teams seamlessly.
- **Reduced remediation latency by 37%** through DynamoDB schema and CRUD service optimizations, eliminating data redundancy and streamlined look ups during root cause analysis.
- Led development of RESTful microservice harnessing .NET, AWS Lambda and API Gateway enabling **real-time user notifications** during remediation with database persistence; **scaling up by 20% user base**.
- **Optimized legacy PowerShell scripts** to start 20+ services concurrently by batching per server; reducing latency from 5 min to 1 min per operation, achieving **70%-time savings** for stakeholders.

Machine Learning Teaching Assistant, Cornell University, Remote, US May 2025 – Aug 2025
- Mentored a cohort of **60+ students** through hands-on labs in **supervised learning, ensemble models, and deep learning**, reinforcing concepts through code reviews, feedback loops, and real-world applications.
- Facilitated **hyperparameter tuning, and feature engineering** in student groups, bridging ML theory with implementation across topics; contributing to a program where **80% of fellows secure internship / job.**