Udayan Atreya

San Jose, CA | 669-340-6033 | udayan.atreya@gmail.com | 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, RAG, HuggingFace, TensorFlow, PyTorch, Git, FastAPI, .NET, RAGAS
LLMs & AI Tools: GPT-4o, Gemini 3 pro, Ollama, Llama, DeepSeek, Groq, Vapi, Eleven Labs
Cloud & Database: AWS, S3, Lambda, API Gateway, CI/CD, MongoDB, DynamoDB, Pinecone, ChromaDB
Languages: Python, Java, C#, SQL

PROJECT EXPERIENCE

Edge-Optimized Agentic RAG Framework *(LangChain, GPT-4o, Gemini 3 Pro, Llama-3, ChromaDB, Whisper)*
Aug 2025 – Present

- Architected an agentic RAG framework implementing Evaluator-Optimizer feedback loop **reducing hallucination by 25%** and **improving answer accuracy by 18%** over baselines for Language-Based Video Reasoning (LBVR) systems.
- Conducted an **ablation study** on embedding architectures, demonstrating that lightweight models (MiniLM) with agentic feedback match heavy SOTA models (BGE-base) at **98% performance parity** while **cutting storage costs by 50%.**
- Engineered an **RAG evaluation suite** using RAGAS and Llama-3-70b (**LLM-as-a-judge**) for 500+ STEM QA pairs **quantifying metrics** like Faithfulness , Answer Relevancy, Response Groundedness, Context Precision, Context Recall, Noise Sensitivity and Semantic Similarity.
- Optimized the latency-accuracy trade-off by implementing Dynamic Chunking and MMR search, reducing **context redundancy by 40%** while maintaining over **90% semantic coverage**.

Agentic BMC Observability Platform *(LangChain, Gemini 1.5, FastAPI, MongoDB, AWS S3)* [GitHub Link]
Jun 2025 – Jul 2025

- Deployed a context-aware agentic observability platform with semantic router, Prometheus and Grafana dashboards for real-time telemetry achieving **95% query classification accuracy** and **reducing manual triage time by 30%**.
- Built a hybrid retrieval engine (MongoDB, S3) synthesizing hot and archived data, accelerating root cause analysis by **cutting retrieval time by 75%** (from 2m to 30s)**.**
- Secured agentic hardware interventions on Baseboard Management Controllers by enforcing strict policy checks, role based and human-in-the-loop validations **preventing 100% of unauthorized state changes**.

WORK EXPERIENCE

Software Engineer, Accenture, Bengaluru, India Dec 2021 – Jul 2024
Associate Software Engineer, Accenture, Bengaluru, India Nov 2020 – Nov 2021

- Architected an user notification system with microservices providing remediation stage notifications using **AWS (API Gateway, Lambda, DynamoDB)** resulting in a **20% increase in DAU** post launch.
- Optimized cloud storage solutions by redesigning **DynamoDB access patterns** (schema, pk, sk) and CRUD microservices **reducing retrieval latency by 37%** ensuring resilient & high performing workloads.
- Delivered a **code vulnerability remediation** solution by using **serverless microservices architecture** with AWS (Lambda, S3, EventBridge) and Azure DevOps**,** safeguarding **400+ applications** across orgs.
- Designed a single touch server patching solution to resolve ServiceNow alerts using serverless architecture (AWS, Azure DevOps), reducing Mean Time to Resolution **(MTTR) by 80%**.
- Partnered with **cross-functional teams** to deliver a **ChatOps platform**, directing natural language queries to automation workflows by integrating serverless microservices, **reducing overhead by 40%** (~5m to ~1m)**.**
- Developed an analytics-driven recommendation engine for 100+ automation assets, implementing weighted scoring logic to personalize tool suggestions for users, **reducing discovery time by 30%.**