# Mukul Rayana

College Park, MD | mukulray@terpmail.umd.edu | +1-667-646-0719
linkedin.com/in/mukul-rayana | github.com/MukulRay1603 | mukulray1603.github.io

## TECHNICAL SKILLS

**Languages:** Python, SQL, C++, C#, Bash
**ML / Deep Learning:** PyTorch, HuggingFace Transformers, Scikit-learn, XGBoost, Reinforcement Learning, ONNX, INT8, CUDA
**GenAI / RAG:** LLMs, Large Language Models, Llama, LangChain, LangGraph, RAG, Agentic AI, LoRA, QLoRA, PEFT, Embeddings, Vector Search, Reranking, FAISS, Pinecone, pgvector
**MLOps / Backend:** FastAPI, REST APIs, Docker, AWS EC2, AWS S3, GitHub Actions CI/CD, MLflow, HuggingFace Spaces, Supabase
**Security / Evaluation:** Prompt Injection Defense, PII Redaction, RBAC (Role-Based Access Control), NLI Guardrails, SHAP, Captum, LLM Evaluation, HITL (Human-in-the-Loop)

## EDUCATION

**University of Maryland, College Park** — College Park, MD
M.S. in Applied Machine Learning — *Expected May 2026*

**SRM Easwari Engineering College** — Chennai, India
B.Tech. in AI and Data Science — GPA 3.55 / 4.0 — 2020–2024

## EXPERIENCE

**BeeBox Studios | IIT Madras Research Park** — Chennai, India
*AI Research Intern* — June 2023 – April 2024

- Built and deployed a BERT-based QA system over a 100-page XR corpus, achieving 83.2% F1 and 71.8% EM for real-time in-headset queries; ONNX export + INT8 quantization reduced P95 latency 3× from 185 ms to 62 ms on GPU-constrained hardware.
- Shipped a Docker + GitHub Actions CI/CD pipeline to AWS EC2 for zero-downtime model updates; collaborated with ML, UI/UX, and WebRTC engineers across the XR prototype stack.

**Campus Recreation | University of Maryland** — College Park, MD
*Operations Supervisor, Climbing & Bouldering Facility* — June 2025 – May 2026

- Promoted to supervisor in 7 months; led 4–6 person shifts, safety briefings, incident response, and new-hire onboarding for a 75+ patron/day facility.

## PROJECTS

**PharmaChain / AI Cargo Monitor** — Apr 2026
*LangGraph, XGBoost, FastAPI, pgvector, React, Supabase*

- Won 1st Place and a $4K team prize at the UMD Smith Agentic AI Challenge; built a LangGraph orchestration of 8 specialized agents over a FastAPI backend for pharmaceutical cold-chain risk triage on simulated telemetry.
- Developed a hybrid risk engine fusing 8 deterministic checks with XGBoost (test ROC-AUC 0.9446) and SHAP explanations; integrated RAG-based GDP/FDA guideline checks over 417 regulatory chunks with human-in-the-loop approval gates.

**EmpathRAG — NLI Safety Guardrail & Emotion-Conditioned RAG** — Apr 2026
*DeBERTa, FAISS, RoBERTa, Captum*

- Fine-tuned DeBERTa-v3-base on 232K NLI pairs to detect crisis-risk language, achieving 0.9629 in-domain recall and 0.75 adversarial recall across 6 attack categories; applied Captum Integrated Gradients for token-level attribution.
- Built a 5-stage emotion-conditioned RAG pipeline over a 1.67M-vector FAISS index with a RoBERTa + LoRA classifier; ablation improved emotion alignment to 0.88 vs. 0.30 BM25 baseline.

**DocPilot — Secure Document Intelligence Platform** — Feb 2026
*RoBERTa, ONNX, BM25, BGE, FastAPI, SQLite*

- Rebuilt an internship-inspired QA architecture into a secure document intelligence platform; engineered prompt-injection defense, PII redaction, RBAC, and append-only SQLite audit logging for auditable document QA.
- Optimized RoBERTa-squad2 QA with ONNX INT8 to 90 ms P95 latency and 66.0% end-to-end F1; used BM25 + BGE dense retrieval with cross-encoder reranking for source-grounded answers.

**RECON — Multi-Agent ML Research Navigator** — *LangGraph, Semantic Scholar, Live* — Mar 2026

- Built a 4-agent LangGraph research assistant with planner, retriever, critic, and synthesizer agents; flagged 52% of outdated ML claims vs. 0% for single-pass RAG on a 130-question eval, while recency-weighted retrieval improved top-paper match from 32.3% to 43.9%.

**Irminsul — LLMOps Fine-Tuning & RAG Serving Stack** — *QLoRA, PEFT, Pinecone, FastAPI, MLflow, Live* — Feb 2026

- Fine-tuned Llama 3.1 8B with QLoRA/PEFT and tracked 3 MLflow experiments, selecting the best checkpoint by semantic similarity 0.826 and ROUGE-L 0.466; deployed FastAPI + Pinecone RAG over 840 docs with weekly GitHub Actions refresh.

## PUBLICATIONS & CERTIFICATIONS

**IEEE IDCIoT 2024:** Voice-Driven Panoramic VR Generation — speech-to-360° pipeline using Whisper, GPT-Neo, and Stable Diffusion; 2.3× throughput via attention slicing on 10 GB VRAM.
**Certifications:** IBM Data Science Professional; UC San Diego Data Structures & Algorithms.