ANIRUD MOHAN

College Park, MD | P: +1 7037058375 | amohan26@umd.edu | LinkedIn | GitHub

SUMMARY

Results-driven AI Software Engineer with a strong foundation in full-stack development, machine learning, and medical AI systems. Experienced in deploying scalable, cloud-based AI applications that integrate multimodal data pipelines and clinical workflows. Proven success in building Retrieval-Augmented Generation (RAG) systems, fine-tuning deep learning models, and implementing advanced imaging solutions for healthcare. Passionate about leveraging AI to improve patient care, diagnostics, and operational efficiency in clinical environments.

EDUCATION

University of Maryland | College Park, MD **Expected Graduation Date - 05/26**
Master of Science in Applied Machine Learning
Coursework: Deep Learning, Computer Vision, Natural Language Processing, Model Optimization, Algorithms for ML, MLOps.

SKILLS

Languages: Python, Java, JavaScript, TypeScript, SQL
Frameworks: FastAPI, React, LangChain, LangGraph, Llama-Index, vLLM
AI/ML Tools: PyTorch, TensorFlow, Hugging Face, MLX, OpenCV, scikit-learn
Cloud & DevOps: GCP, Docker, Kubernetes, Git, CI/CD, Azure DevOps
Data Systems: PostgreSQL, MongoDB, Databricks, Snowflake
Specialized Areas: RAG systems, LLM orchestration, Medical Imaging (MRI, Diffusion Models), OCR & Data Extraction
Security & Testing: OAuth2, JWT, Unit Testing, Secure Coding Standards

WORK EXPERIENCE

CarinaAI **Herndon, USA**
Data Science Intern Oct 2025 – Dec 2025

- Engineered an automated Chart Review RAG pipeline for the **DeIdentifier** product, boosting system accuracy from **69% to 82%** through advanced post-processing and output flow validation.
- Optimized retrieval performance by benchmarking diverse embedding models, implementing **query transformation**, and refining **reranking strategies** to enhance context precision.
- Orchestrated the deployment cycle using the **vLLM framework** and **KV caching** to minimize memory footprint, while securing source code via **Cythonization** for client-side distribution.
- Diagnosed and resolved a critical "Yes/No" evaluation bias by validating model outputs against physician-curated ground truth, correcting the evaluation schema and improving reliability.

Thapovan Info Systems **Chennai, IND**
Junior Machine Learning Engineer Oct 2023 – Jun 2024

- Engineered an **Advanced Retrieval-Augmented Generation (RAG) chatbot** trained on medical insurance data, enabling accurate, context-aware responses for policy and claim inquiries.
- Built a **data ingestion and retrieval pipeline** from scratch integrating OCR-based PDF parsing, text normalization, and structured chunking for high-recall document search.
- Designed **guardrail systems** to improve retrieval accuracy and response reliability, reducing irrelevant model outputs by 40%.
- Experimented with **LoRA fine-tuning on the MLX framework**, later optimizing performance and cost by pivoting to RAG-based strategies.
- Collaborated with data and product teams to deploy scalable AI services via **FastAPI** and **Dockerized microservices** integrated with cloud environments.

Azentio Software **Chennai, IND**
Software Developer Intern Feb 2023 – Sept 2023

- Refactored a monolithic application into microservices; implemented REST APIs and caching mechanisms.
- Built Python-based data extraction and transformation pipelines (Scrapy, BeautifulSoup).
- Collaborated on system performance improvements and provided integration testing for scalable deployments.

COURSEWORK & CAPSTONE PROJECTS

Pathology-Controllable Diffusion for Brain MRI | *Sep 2025 – Dec 2025*

- Led dataset engineering by curating 5-class segmentation masks (GM, WM, CSF) using ANTs Atropos, integrating healthy tissue labels to explicitly supervise the model's understanding of anatomical preservation.
- Engineered a scalable training pipeline in PyTorch/Diffusers, implementing Automatic Mixed Precision (AMP) to optimize throughput on A100 GPUs and establishing MLOps protocols for real-time SSIM tracking via Weights & Biases.
- Benchmarked counterfactual quality across novel metrics (Tumor Residual, DiffMap IoU) and standard baselines (FID), achieving a 66% reduction in tumor abnormality and a best-in-class FID of 58.1

AI-Powered Code Reviewer — MLOps Driven Solution | *Feb 2025 – May 2025*

- Fine-tuned Microsoft's CodeReviewer using LoRA + DDP on 150K+ GitHub PRs, achieving BLEU 80.53 and EM 26%, surpassing SOTA baselines.
- Deployed inference APIs with FastAPI + Docker, hosted on Hugging Face Spaces for scalable, low-latency serving.
- Automated model integration via GitHub Actions CI/CD, reducing developer feedback cycles by 15%.