Khaja Moinuddin Shaik Mohammed

+91 9908152590 | monu.dev@proton.me | github.com/codewithmoin | linkedin.com/in/codewithmoin

Applied Scientist — LLM Systems · NLP · Machine Learning

Summary

Applied Scientist Intern and Graduate CS (AI/ML) student with research and industry experience in LLM systems, NLP, and on-device ML. Interned at Amazon RBS Sciences building autonomous taxonomy generation pipelines; first author at AMLC 2026. Focused on building scalable, production-ready AI systems from retrieval infrastructure to model optimization.

Education

Mahatma Gandhi Institute of Technology Oct 2022 – Jul 2026 (Expected)

B.Tech in Computer Science (Artificial Intelligence and Machine Learning) *Hyderabad, India*

- **GPA: 9.09/10.0**

Experience

Applied Scientist Intern — Amazon Feb 2026 – Jun 2026

RBS Sciences *Bangalore, India*

- Built an autonomous LLM-powered taxonomy generation platform that transformed millions of customer feedback records into hierarchical taxonomies, enabling scalable issue discovery across Amazon businesses while outperforming the manual baseline (0.74 vs. 0.71 F1).
- Designed a self-calibrating Knowledge Information Extraction pipeline that eliminated manual prompt engineering, reducing new business onboarding from 5–7 days to under 18 hours.
- Developed an explainable taxonomy evaluation framework enabling rapid taxonomy validation through grounded, human-readable quality diagnostics.

AI Intern — Intel Corporation May 2025 – Jul 2025

Intel Unnati Industrial Training Program

- Built a **real-time video feed sharpening pipeline** using **knowledge distillation**, improving edge clarity by 20% and reducing model size by 30% for deployment on resource-constrained devices.
- Optimized CNN inference latency by 35% on low-resource devices using PyTorch, enabling smooth live-feed processing without dedicated GPU hardware.

Selected Projects

DocuLens AI | *LLMs, RAG, FastAPI, PostgreSQL, pgvector, Redis, OpenAI* Jun 2025 – Present

- Built an enterprise document intelligence platform enabling semantic search, summarization, classification, and question answering over 10K+ unstructured documents using Retrieval-Augmented Generation.
- Engineered a low-latency RAG pipeline achieving sub-800ms retrieval and under 3s end-to-end response time using OpenAI embeddings, pgvector, and Redis caching.

EcoGuardian AI | *TensorFlow Lite, React Native, Python, Appwrite* Mar 2025 - Apr 2025

- Built an on-device waste classification system achieving 82% accuracy with sub-300ms inference using quantized TensorFlow Lite models.
- Reduced model size by 60% through post-training quantization, enabling fully offline deployment without cloud dependency.

Technical Skills

Languages: Python, SQL, C++, Java, TypeScript
Machine Learning: PyTorch, TensorFlow, Scikit-learn, Transformers, BERTopic, HDBSCAN, UMAP
Generative AI: Context Engineering, RAG, LLM Evaluation, Amazon Bedrock, OpenAI API
ML Infrastructure: FastAPI, Docker, Redis, pgvector, Prompt Caching, Distributed Inference
Cloud & Tools: AWS (Bedrock, SageMaker, S3, EC2), GCP, Git, Weights & Biases

Achievements

Universal Anecdote Miner (UAM) — First Author 2026

Amazon ML Conference (AMLC) 2026 Submission.

LUMEN — Third Author 2026

AMLC 2026 Submission · EMNLP 2026 Submission.

Google Student Ambassador Aug 2025 – June 2026

Represent Google's developer community at MGIT.