

Khaja Moinuddin Shaik Mohammed

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

AI Engineer | LLM Agents · RAG · Document AI · Production AI Systems

Hyderabad, India

SUMMARY

AI Engineer building LLM, retrieval, and agentic systems in Python. At Amazon, built autonomous LLM workflows over millions of customer records and an evaluation framework adopted by 12 business teams. Experienced in multi-agent orchestration, document retrieval, evals, backend APIs, async pipelines, and production reliability.

EXPERIENCE

Applied Scientist Intern — Amazon

Jan 2026 – Jun 2026

RBS Sciences

Bangalore, India

- Built an autonomous LLM pipeline over millions of customer feedback records to generate hierarchical issue taxonomies, outperforming the manual scientist baseline with **0.74 vs. 0.71 F1**.
- Built an explainable **LLM evaluation framework** with human-readable quality diagnostics, adopted by **12 Amazon business teams** to evaluate and validate generated taxonomies.
- Designed a self-calibrating extraction pipeline that removed per-domain prompt engineering, reducing new-business onboarding from **5–7 days to under 18 hours**.

AI Intern — Intel Corporation

May 2025 – Jul 2025

Intel Unnati Industrial Training Program

Hyderabad, India

- Compressed a real-time video model using **knowledge distillation and quantization in PyTorch**, reducing model size by 30% and inference latency by 35% on GPU-free edge devices while improving clarity by 20%.

PROJECTS

Attest — agentic system for grounded, self-verifying answers | *Python, LangGraph, FastAPI, Postgres, Redis* 2026

- Built an agentic RAG workflow that verifies each generated claim against the **exact retrieved evidence**, using an independent judge model and bounded revision before escalating uncertain answers.
- Implemented hybrid retrieval with **pgvector + PostgreSQL full-text search**, reranking, and Redis-cached LLM calls for grounded document question answering.
- On a 41-question benchmark, verification reduced failing citations from **9.87% to 4.15%**; achieved **0.473 evidence F1** on human-annotated Qasper.
- Designed explicit escalation and evaluation metrics to distinguish correct answers, unsupported claims, and cases requiring human review.

Decode — multi-agent production workflow with artifact lineage | *Python, FastAPI, Redis, ARQ, PostgreSQL* 2026

- Built an **8-department multi-agent orchestration system**, with each agent running bounded generate/evaluate/revise loops and escalating to a human after failed retries.
- Implemented versioned artifacts, content hashes, and a dependency graph so outputs remain traceable to their inputs and only affected downstream work is regenerated.
- Designed production reliability primitives including a **transactional outbox**, idempotency keys, advisory locks, stale-run fencing, and resumable SSE for observable, recoverable async execution.

RESEARCH

Universal Anecdote Miner (UAM) — First Author, AMLC 2026 submission **LUMEN** — Third Author, AMLC 2026 & EMNLP 2026 submissions

EDUCATION

Mahatma Gandhi Institute of Technology

Oct 2022 – Jul 2026

B.Tech, Computer Science (AI and Machine Learning)

Hyderabad, India

- GPA 9.09/10.0 · Amazon ML Summer School 2025** — selected 3,000 of 165,000+ applicants (~top 2%).

TECHNICAL SKILLS

Languages: Python, SQL, TypeScript, C++, Java

LLMs & Agents: LLMs, RAG, AI agents, LangGraph, multi-agent orchestration, LLM-as-judge, context & harness engineering, agent evaluation, benchmark design

Machine Learning: PyTorch, Transformers, Deep Learning, NLP, model optimization & quantization

Backend & Infrastructure: FastAPI, PostgreSQL (pgvector, FTS), Redis, ARQ, Docker, AWS (Bedrock, SageMaker, S3, EC2), GCP, Git, Langfuse