MAJID HUSSAIN

AI Engineer

Lahore, Pakistan | 0305-8339832 | majidhussainf543@gmail.com | linkedin.com/in/majidhussain7 | github.com/Majid54H

SUMMARY

AI Engineer with hands-on experience building RAG pipelines, integrating LLMs, and developing real-time computer vision systems. Promoted from intern to full-time at TechReFi within 3 months for delivering AI features end-to-end, from model integration to deployment on AWS. Comfortable across the ML stack: embeddings, vector search with Qdrant, CPU-optimized inference with OpenVINO, and REST API design.

TECHNICAL SKILLS

Languages & Backend: Python, FastAPI, REST API, C#, C++, Node.js (TypeScript), Angular
LLMs & NLP: Llama 2/3, Mistral 7B, GPT-4, GPT-4o, Claude, Hugging Face Transformers, Sentence-Transformers (all-MiniLM, BGE), spaCy
RAG & Vector DBs: Qdrant (HNSW), FAISS, Hybrid Search, Cross-Encoder Reranking
Computer Vision: YOLOv8/v9, RT-DETR, OpenCV, DeepSORT, ByteTrack, OpenVINO (INT8/IR)
Cloud & MLOps: AWS (EC2, S3), Docker, Git, MLflow, JWT Auth

EXPERIENCE

AI Engineer · TechReFi *Sep 2025 – Present | Lahore, Pakistan*

Intern → Full-Time (promoted within 3 months) · On-site

- Built a RAG platform that indexes company documents (manuals, technical docs, procurement records) into a Qdrant vector database.
- Used semantic chunking and hybrid search so employees can search across all documents in natural language.
- Deployed the platform on AWS EC2 with FastAPI; it now powers automated RFP/RFQ extraction and vendor matching.
- Combined metadata filtering with cross-encoder reranking to cut search time to under 200ms.
- Reduced unnecessary full-database scans by narrowing the set of results sent to reranking.

AI/ML Engineer Intern · CodeCelix *Aug 2025 – Nov 2025 | Remote*

- Built and integrated LLM-powered projects, including a RAG chatbot (Qdrant, GPT-4) that automated common support queries.
- Fine-tuned an open-source GPT-3 style model to generate image captions and hashtags from a captions dataset.
- Led a team of 4 interns to build a CV parsing pipeline using spaCy for entity extraction and BERT-based classifiers.
- Improved the recruiter's candidate shortlisting workflow by automating resume screening.

PROJECTS

Real-Time CCTV Analytics System — TechReFi

Python · YOLOv8/v9 · OpenVINO · OpenCV · Intel CPU

- Built a real-time CCTV analytics tool for footfall counting, demographic detection, and attendance tracking for a retail client.
- Switched inference from GPU to OpenVINO INT8 quantization on Intel CPU, cutting CPU load per stream from ~60% to ~15%.
- Sustained 30+ FPS across 8 camera streams running at the same time.
- Designed one shared inference process for all cameras, with a separate capture process per camera.
- Added a watchdog that automatically recovers dropped camera streams, keeping the system stable in production.

AI Procurement Analysis & Automation Platform — TechReFi

Python · GPT-4o · RAG · Qdrant · FastAPI · AWS EC2

- Built LLM pipelines to extract and analyze RFP/RFQ documents for vendor evaluation.
- Cut manual document review time by automating requirement parsing with multi-step prompts.
- Built a reranking layer (cross-encoder scoring over Qdrant results) to compare vendors on price, quality, and compliance.
- Mapped the extracted output into ERP-compatible data formats.

Multi-Modal RAG Document Intelligence System — TechReFi

Python · CLIP · BGE-Reranker · Qdrant · FastAPI · AWS

- Built a system to ingest and search PDFs, images, and structured documents together.
- Used CLIP for image embeddings and a BGE cross-encoder to rerank search results.
- Applied semantic chunking and query-aware context building to improve retrieval accuracy.
- Added automatic query rewriting to return more relevant results.
- Extracted structured metadata to support knowledge-graph-style retrieval.
- Built a secure WebSocket API to support multiple concurrent file uploads.

EDUCATION

B.S. Software Engineering · Lahore Garrison University

Nov 2021 – Jun 2025 — Machine Learning (A) · Data Structures & Algorithms · Computer Vision · OOP