MOHAMMED ZAIN RAFEEQUE

Email: zainrafeeque@gmail.com |**Contact Number:** +971-56 957 0799

LinkedIn: linkedin.com/in/zain-rafeeque/ | **GitHub**:github.com/ZainRafeeque

Summary

AI Engineer with hands-on production experience designing and deploying intelligent systems across the UAE. Currently building agentic RAG pipelines, multi-modal document intelligence, and domain-specific LLM applications at MTA Investment LLC in Dubai. Proven track record delivering end-to-end AI solutions — from a fine-tuned bilingual Glass Industry Expert System with hybrid vector search and 247K+ chunk retrieval, to workflow automation and scalable API integration. Proficient in Python, LangChain, FastAPI, pgvector, vLLM, and n8n, with deep expertise in LLMs, retrieval-augmented generation, and real-world ML deployment. Experienced in working with leading AI platforms and APIs including Claude (Anthropic), Gemini Vision, OpenAI, Groq, and Manus AI, with a strong ability to integrate and orchestrate multiple AI services into production-ready systems. Recognized for excellence in AI prompt engineering by the Dubai Centre for AI (DCAI).

Education

Dec 2020 – Sep 2024

Bachelor of Technology, Artificial Intelligence, and Data Science

Vimal Jyothi Engineering College, approved by the All-India Council for Technical Education (AICTE) and affiliated with the APJ Abdul Kalam Technological University

Experience

Oct 2025 – Current

AI Engineer – MTA Investment LLC – Dubai:

- Glass Engineer — Bilingual RAG Expert System: Architected a production-grade bilingual (English + Farsi) AI assistant for glass manufacturing engineers using a fine-tuned Qwen2.5-14B LLM (vLLM, 4-bit LoRA), hybrid retrieval (pgvector + BM25 + Reciprocal Rank Fusion) over 247K+ chunks, cross-encoder reranking, semantic caching (Redis), and a React 18 + TypeScript chat UI — deployed via Docker Compose achieving ~85% retrieval hit rate and 1–5s end-to-end latency.

- Engineering Agentic RAG Systems: Architecting production-ready autonomous retrieval using ChromaDB, MySQL, and Google Drive for intelligent data querying.

- Advanced Workflow Automation & Multi-Modal Document Intelligence: Built end-to-end operational flows using n8n to streamline complex business processes, deployed Gemini Vision API via Manus AI to automate data extraction from multi-lingual scanned PDFs, and developed custom APIs to connect Manus AI with internal databases, simplifying AI access for non-technical users.

Sept 2025 – Oct 2025

AI Freelance – Khansaheb Sustainability – Dubai:

Architecting automated workflows using n8n to streamline operations, alongside developing a RAG + LLM chatbot prototype via LangChain and Groq API to automate customer inquiries, deliver smart product advice, and optimize support efficiency.

May 2025 – Aug 2025

AI Intern – Direct Axis Technologies – Dubai:

Deployed real-world AI and ML solutions, such as an intelligent cheque processing system powered by neural networks and Python, to automate data extraction, validation, and financial workflows.

Oct 2024 – Mar 2025

AI/ML Trainee - Meta Scifor Technologies:

- Developed expertise in machine learning algorithms and deep learning frameworks while working on different projects.
- Took part in a comprehensive remote training program to build proficiency in Python programming.

May 2023

Python Intern, ReverTech:

- Enhanced Python programming skills through a practical internship at ReverTech, with a focus on data structures and algorithms.

Projects

Mar 2026

AI-Powered Glass Industry Expert System (RAG + vLLM + pgvector + React)

- Architected a production-grade bilingual (English + Farsi) AI assistant for glass manufacturing engineers, leveraging a fine-tuned Qwen2.5-14B LLM (4-bit LoRA, vLLM) with a 4-stage hybrid RAG pipeline combining pgvector HNSW dense search over 247K+ chunks, BM25 keyword search, Reciprocal Rank Fusion, and cross-encoder reranking — achieving 85% retrieval hit rate.
- Deployed full stack via Docker Compose with a React 18 + TypeScript chat UI, JWT auth, Redis semantic caching, and specialized endpoints for glass composition analysis, formulation design, and defect troubleshooting.

Jan 2026

Agentic RAG System — Multi-Source Data Ingestion & Retrieval Pipeline

- Architected a production-ready autonomous RAG system integrating multiple data sources — local files (PDF, DOCX, TXT, CSV), PostgreSQL/MySQL, REST APIs, and Google Drive — via a ConnectorFactory pattern, with a dedicated IngestionAgent orchestrating extract → chunk → embed → store workflows into ChromaDB.
- Implemented sentence-boundary chunking with three vectorization strategies (pre-vectorize, on-demand, hybrid) and per-client YAML/JSON config support, enabling flexible and scalable document ingestion for intelligent querying across internal business systems.

Sep 2025

AI-Powered Product Assistant Chatbot (RAG + LangChain + Groq API)

- Engineered an intelligent product assistant chatbot for Khansaheb Sustainability, leveraging Retrieval-Augmented Generation (RAG) and LangChain with Groq API LLM to handle product-related queries and enhance customer support.
- Enabled real-time, context-aware responses by integrating dynamic document retrieval and conversational memory, streamlining sustainability product inquiries.

April 2024 **A Novel Approach to Malayalam Speech-to-Text and Text-to-English Translation**

• Pioneered a solution to facilitate seamless communication between Malayalam and English speakers through advanced NLP techniques, resulting in a 40% increase in translation efficiency.

• Significantly improved translation accuracy and speed by leveraging state-of the art language models.

Skills

- **Programming Languages:** Python, R, Java, SQL
- **Data Science Libraries & Frameworks:** NumPy, Pandas, Scikit-learn, TensorFlow/PyTorch, Keras, Spark
- **AI/LLM Tools:** LangChain, RAG, vLLM, OpenAI API, Groq API, Gemini Vision API, n8n, ChromaDB, pgvector
- **Data Visualization:** Matplotlib, Seaborn, Plotly
- **DevOps & Infrastructure**: Docker, FastAPI, PostgreSQL, Redis, Nginx, Git

Certification

- **One Million Prompters (AI Prompt Engineering)** - Dubai Centre for Artificial Intelligence (DCAI) | 2025
- **Data Analytics Assessment** by LearnTube (Career Ninja) | 2024

Courses

- **Deep Learning - A real-world approach add-on course:**

Enhanced Deep Learning Skills, under the mentorship of a Microsoft Research Scientist, mastered the art of building and training neural networks.

- **Linux101.1x: Shell Programming:**

Participated in an IIT Bombay course to master shell scripting for automation and system administration.

- **Python101x: Basic Programming using Python:**

Participated in an IIT Bombay course to master Python programming concepts for data science and software development.

Personal Details

- **Nationality** Indian
- **Visa Status** UAE Resident