Ali Raza

+92 300 7651218 mirzaalirazafsd@gmail.com ali-raza-0a1282281 alyrraza alyrraza.vercel.app

Summary

AI Engineer with 2 years of experience building agentic and RAG applications for US clients, from ambiguous requirements to production. Technical lead on a live AI SaaS. Builds evaluation with LLM judges, frozen baselines and paired bootstrap tests, and designs safe fallback paths so agents stay reliable.

Technical Skills

ML / DL & CV: PyTorch, TensorFlow, Scikit-learn, HuggingFace Transformers, **Optimization**(LoRA, QLoRA, PEFT) , **ML Models**(XGBoost, LightGBM), **Computer Vision**(YOLOv8, U-Net, ViT, SAM, BioViL-T, OpenCV), **Deep Generative Models**(GANs, VAEs, Diffusion)

NLP, LLMs & Data: BERT, LlamaIndex, LangGraph, RAG pipelines, agentic workflows, LangChain, vector databases, FAISS, pgvector, embeddings, prompt engineering, **LLM-as-judge evaluation**, MCP, OpenAI, Gemini, Anthropic APIs, Pandas, SQL, PostgreSQL

MLOps, Deployment & Programming: MLflow, Docker, Kubernetes, Kubeflow, Prometheus, Grafana, FastAPI, REST APIs, Redis, Celery, CI/CD, GitHub Actions, GHCR, AWS EC2, inference optimization, GPU deployment, Python, JavaScript, Git

Experience

AI Engineer — HusQuay Solutions 2025 – ongoing
Remote, Faisalabad, Pakistan

- Integrated open-source computer vision models with third-party **VLM and LLM APIs** into a unified inference pipeline for a production SaaS platform, replacing a planned GPU retraining system with a faster API-driven approach.
- Owned the end-to-end architecture from raw image input to report-ready structured output (supporting classification and detection models feeding LLM/VLM classification), automating property tax depreciation studies for US clients in a live system.
- Maintained **MLflow** experiment tracking and model versioning; automated deployments via **CI/CD** pipelines using GitHub Actions.

Data Analyst Intern | TAKMIL 2024
Louisville, Kentucky, USA · Remote

- Analyzed large datasets (Python, SQL); built **Power BI dashboards**; proposed optimizations improving workflow efficiency by **30%**.

Research & Key Projects

DocDesk: Healthcare RAG Service with an Evaluation Harness GitHub 2026
FastAPI · PostgreSQL + pgvector · Redis · Celery · LLM-as-judge · Docker · GitHub Actions · GHCR

- Built a RAG service over 298 MedlinePlus and DailyMed documents with citations, **abstain and refuse** paths and a web UI; image published to GHCR from CI.
- Engineered evaluation beyond accuracy: LLM judges (faithfulness, correctness), citation precision and recall, failure labeler, CI retrieval gate; prompts, judges and configs **hash-pinned**; response cache with a token budget guard re-ran 44 items at **0 tokens billed**.
- Ran 7 experiments against a frozen baseline with paired bootstrap intervals: chunk cap lifted strict pass **15/30 to 19/30** on dev, 13/20 to 15/20 on held-out test (not significant); an independent review found **8 major defects**, fixed first.

AI Cost Segregation Platform — Multimodal VLM Pipeline, Production SaaS 2025–Present
Gemini VLM · YOLOE · SAM2 · FastAPI · Firebase · Prompt Engineering

- Technical lead on an AI SaaS converting property photos into audit-ready tax reports; replaced a fine-tuned closed-vocab detector with a **Gemini VLM classification stage** (312-item taxonomy), eliminating the retraining loop.
- Designed a **3-stage vision pipeline (YOLOE + SAM2 + Gemini)** with IoU deduplication.

CV Interview Coach — AI Multi-Agent System 2025
LangGraph · LangChain · Groq · ChromaDB · LlamaIndex · Whisper · LiveKit · FastAPI · Next.js

- Built a **3-agent pipeline** with a supervisor orchestrator: CV/JD analyzer, RAG-augmented question generator, and real-time voice interview agent with **self-healing retry loops** (critique injection, max 3 retries).

Voice-Driven B2B Lead Generation Agent — AI/MLOps 2025
LiveKit · Google STT/TTS · Tavus · n8n · MCP · aiohttp · Webhooks

- Built a real-time voice agent on LiveKit (Google STT/TTS, Tavus) that collects, normalizes and confirms B2B lead data in a four-step conversation, with asynchronous n8n webhooks delivering results by email after confirmation.

RadGuard: Multimodal ML System for Medical Imaging — Final Year Thesis 2025–2026
BioViL-T · CXR-BERT · PyTorch · MLflow · Docker · Kubernetes · Kubeflow · FastAPI · AWS EC2

- Built a multimodal system detecting 4 error types in AI-generated chest X-ray reports across 14 conditions on 74,060 MIMIC-CXR samples (**Val F1=0.66**); fixed 10 critical labeling bugs (147K false labels); deployed on **AWS EC2** with Docker, Kubernetes autoscaling, **Kubeflow** retraining, Prometheus+Grafana monitoring and a **FAISS RAG** chatbot.

Education

BSc Computer Science | FAST-NUCES, Faisalabad 2022 – 2026
Co-Head Hackathon ACM Daira 2025 | TAKMIL Data Science Internship 2024

Publications

Khan MM, **Raza A**, et al. "Generalizable Detection of CIEDs Using a Hybrid Transformer and CNN Framework" — **ACR 2026 [Accepted]**

Ali Raza. "Learn2Breath: Deformable Lung CT Registration" — **Learn2Reg / MICCAI 2026 [Ranked 3rd, Invited Oral]**