

MUHAMMAD AUN ALI

+92 333 0226005 • mohammadaliaun7@gmail.com • muhammadaun.dev
linkedin.com/in/muhammad-aunali • github.com/SeekAI-786

SUMMARY

AI Engineer with hands on experience fine-tuning computer vision models (YOLO, RF-DETR) and building RAG/LLM powered applications, from data annotation and model training through deployment. Shipped full stack AI systems spanning video intelligence, natural language analytics and conversational AI.

WORK EXPERIENCE

AI Engineer , Liquid Technologies

Nov 2025 - June 2026

- Fine-tuned YOLO & RF-DETR models for real-time detection (PPE, fire, shoplifting, jewelry; up to 10+ classes) on 2,000-3,000 annotated images, achieving ~40 mAP at 30-40 FPS.
- Built 2-stage face recognition pipeline (RetinaFace + ArcFace) integrated into live video analytics via REST API and MongoDB, cutting inference cost 50%+ with an in house model
- QA and integration testing on Sally, an AI business assistant, validating response flows using Playwright and Selenium; caught a critical OTP verification bug
- Developed action recognition system using CNN+LSTM and TSM for shoplifting, fall and slip detection.

AI Intern, Trust Nexus

July 2025 - Oct 2025

- Deployed real-time emotion, age and gender detection at ~25-30 FPS, achieving ~80% accuracy on gender classification and ~60% on age estimation for NEXI Humanoiad Robot.
 - Built a biometric verification system using ArcFace and Resemblyzer embeddings for speaker recognition, achieving ~85% accuracy on internal test set.
-

PROJECTS

SmartBiz - DA Assistant, FYP Project

Oct 2025 - June 2026

- Engineered SmartBiz – DA Assistant end to end (React, FastAPI, DuckDB); NL to SQL layer resolved 90% of queries correctly at ~1.7-2s avg response time
- Built dual LLM pipeline (OpenAI primary, Gemini fallback) with read only SQL validation and prompt injection guards; auto-generated BI reports (~19s) and dashboards (~6s)

Pak Constitution ChatBot, Freelance Project

Jan 2026 - Feb 2026

- Built a RAG based Q&A system indexing 212 pages of legal documents via LangChain, Google Gemini and Pinecone, with average query response time of 5-10 seconds.

MediBot - CustomGPT

Apr 2025 - May 2025

- Fine tuned a custom LLM on 50,000+ medical Q&A pairs using Python; improved response accuracy by 35% over baseline through iterative prompt and training optimization.
-

EDUCATION

Bachelor of Science in Artificial Intelligence
SZABIST University Karachi

Sep 2022 - July 2026

SKILLS

- **Programming:** Python, SQL
- **ML/DL Libraries:** PyTorch, TensorFlow, Keras, Scikit-Learn, Transformers
- **Computer Vision:** OpenCV, YOLO, RF-DETR, DeepFace
- **GenAI & RAG:** LangChain, Pinecone, Google Gemini, Vector Embeddings
- **Data & Visualization:** Numpy, Pandas, Matplotlib
- **Backend & APIs:** FastAPI, REST APIs, MongoDB
- **Tools:** Git, GitHub, Docker, Vercel