

# MUHAMMAD QASIM ALI

AI/ML Engineer · MLOps & LLMOps · RAG / LLM Systems  
03366909047 · alimqasim427@gmail.com · Lahore, Pakistan  
[Linkedin](#) · [Github](#) · [Portfolio](#)

## SUMMARY

AI/ML Engineer focused on Computer Vision, Machine Learning, and Generative AI systems. Experienced in building RAG pipelines, deploying ML models with MLOps practices, and optimizing LLM inference for edge devices. Skilled in Python and deep learning frameworks, with a focus on scalable, production-ready AI systems.

## TECHNICAL SKILLS

**Languages:** Python, C++, SQL, Bash

**Machine Learning & Deep Learning:** PyTorch, TensorFlow, OpenCV, CNNs, LSTM, U-Net, ResNet-50, YOLOv12, Grad-CAM, image segmentation, object detection, facial recognition, computer vision, model training & evaluation

**LLMs & Agentic AI:** Retrieval-Augmented Generation (RAG), LangChain, vLLM, Milvus & FAISS (vector databases), hybrid search (BM25 + dense embeddings), BGE embeddings, reranking, prompt engineering, n8n (agent workflows), LLaMA 3, Qwen

**MLOps & Deployment:** Docker, MLflow (experiment tracking & model versioning), FastAPI, Flask, REST APIs, Git & GitHub, GitHub Actions (CI/CD), model serving & inference optimization, Linux

**Edge AI:** NVIDIA Jetson (Orin Nano), Raspberry Pi, on-device / embedded AI

## EXPERIENCE

### Punjab Information Technology Board (PITB) — AI Intern

Jun 2025 – Aug 2025

- Engineered a U-Net semantic-segmentation pipeline predicting soil-moisture index masks from 13-band Sentinel-2 multispectral imagery (10–60m resolution), achieving 0.89 mean IoU, with MLflow for experiment tracking, model versioning, and performance monitoring.
- Deployed a hybrid-search RAG chatbot for the “Maryam Ki Dastak” program using BM25 + Milvus (BGE embeddings) and Qwen3-8B-AWQ served via vLLM, with hallucination guardrails and reranking, improving retrieval accuracy by 22% and cutting irrelevant responses by 30%.

### Nayatel — AI Intern

Jul 2024 – Aug 2024

- Built a customer-support chatbot with LangChain RAG pipelines and LLaMA 3 (8B) on a domain-specific 5k Q&A dataset, achieving 92% precision in human testing.
- Deployed it via Flask + ReactJS and containerized with Docker, handling 1,000+ daily queries at low latency.

## PROJECTS

### Multimodal Humanoid Robot for Autistic Children *Final-Year Project*

- Built an audio-video companion robot on a dual Raspberry Pi architecture, integrating real-time object detection, facial recognition, and voice-command processing to improve communication and engagement for autistic children.

### AI-Powered PCOS Assistant *Healthcare AI · Multimodal*

- Engineered a multimodal PCOS pipeline pairing a ResNet-50 model that detects follicular patterns in ultrasound scans (95%+ accuracy, with Grad-CAM interpretability) and a LangChain RAG symptom analyzer, plus an n8n agent for automated appointment scheduling.

### Cotton Detection & Yield Prediction *Computer Vision · AgriTech*

- Built a YOLOv12 cotton-detection system trained on ground-level imagery from multiple farm locations, reaching mAP@0.5 of 94% for plant detection to support yield prediction.

### Urdu Audio Deepfake Detection *Audio · Deep Learning*

- Developed a CNN + LSTM model on Mel-spectrogram and MFCC features to detect AI voice-cloning in Urdu, achieving 92% accuracy and addressing misinformation in a low-resource language.

## EDUCATION

### Ghulam Ishaq Khan (GIK) Institute of Engineering Sciences and Technology

Sep 2022 – Jul 2026

B.S. in Artificial Intelligence — Topi, Pakistan

Relevant Coursework: Data Structures & Algorithms, Machine Learning, Deep Learning, Computer Vision

## LEADERSHIP

### Team Techno (GIKI Robotics) — Team Manager

Sep 2025 – May 2026

- Led a 45+ member robotics team; improved technical onboarding for 25+ new members and secured top placements at major national competitions.

### Microsoft Club GIKI — Partnership & Outreach Lead

Nov 2024 – Apr 2025

- Designed tiered sponsorship proposals and maintained sponsor relationships through regular project updates.

## HONORS & PUBLICATIONS

- 1st Place** — All Pakistan Prompt Engineering Competition (2025).
- Publication:** “Malaria Cell Classification Using CNN: A Deep Learning Approach,” ICAI 2024, Varna, Bulgaria. [DOI](#)
- Hackathons:** Co-Creating with GPT-5, ElevenLabs Voice AI, and Llama 2 × Clarifai, NextGen-GPT