

OMAR BAIG

Lahore, Pakistan | omarbaig8070@gmail.com | +92 334 0101065
linkedin.com/in/omar-baig-b14a372b1 | github.com/OmarBaig05 | omarbaig.dev

SUMMARY

MITACS Globalink-funded AI/ML Engineer and Researcher specializing in **agentic LLM systems, retrieval-augmented generation, and multimodal computer vision**. Shipped 10+ production AI systems and a first-author manuscript on hierarchical preference graphs and federated LoRA adaptation for personalized LLM agents; designed a novel Cross-View Attention architecture achieving 83% AUC across 14 respiratory pathologies. Combines end-to-end ownership — from model architecture to full-stack deployment — with a track record of building explainable, privacy-preserving AI for healthcare and clinical decision support.

EDUCATION

University of Engineering and Technology (UET), Lahore Dec 2022 – Jun 2026
B.S. in Computer Science CGPA: 3.56 Lahore, Pakistan

- **Thesis:** AiroDx — Context-Aware Explainable AI for Multimodal Differential Diagnosis of Respiratory Diseases.
- **Relevant Coursework:** Artificial Intelligence, Data Structures & Algorithms, Database Systems

EXPERIENCE

MITACS Globalink Research Intern Jul 2026 – Present
University of Winnipeg Winnipeg, Manitoba, Canada

- Architected a full-stack, LLM-integrated research application for contextualizing health data, enabling faster interpretation of complex patient information for a next-generation smartwatch health platform.
- Own application architecture end-to-end, leading a 3-person research team (1 faculty supervisor, 2 undergraduate researchers) to translate open-ended research goals into shippable software components.
- Built TypeScript-based frontend and backend services, REST API integrations, and database-backed data flows spanning the full application stack.
- Designed reusable LLM integration and retrieval abstractions for research experimentation, prioritizing maintainable architecture and reliable contextual processing.

Freelance AI & Machine Learning Engineer Oct 2025 – Present
Fiverr (Level 1 Seller) Remote

- Built and shipped a token-gated AI SaaS chatbot (FastAPI, React/TypeScript, MongoDB) with RAG-based document retrieval, Stripe subscription billing, Clerk authentication, and an analytics admin dashboard, as one of 10+ client AI/ML projects delivered with a 5-star rating.
- Engineered Groq-backed LLM microservices with vector search, rate limiting, and sentiment analysis, deploying a 4-tier (guest/free/pro/premium) query-enforcement system to a production VPS.
- Trained and optimized ML models across diverse client datasets, delivering adaptive LLM-integrated query systems and automated decision-support tools for real-world business use cases.

Generative AI Intern Jun 2024 – Aug 2024
Xavor Corporation Lahore, Pakistan

- Led a cross-functional developer-clinician team to build an intelligent mobility-assessment chatbot automating Functional Reach Tests, reaching 95% accuracy using OpenCV and MediaPipe pose estimation.
- Engineered a real-time pose-estimation pipeline achieving 90% validated measurement accuracy via geometric modeling techniques.
- Ranked 1st among 10+ competing internship teams for technical execution and clinical impact.

PROJECTS

AiroDx — Multimodal Respiratory Diagnosis | *PyTorch, timm, Groq* [1st Place FYP] Oct 2025 – May 2026

- Achieved 83% AUC across 14 respiratory pathologies by architecting a novel Multi-View DINO-ViT model with **Cross-View Attention** to fuse frontal and lateral chest X-ray features, mimicking radiologist reasoning across orthogonal projections.
- Built a dual-backbone Vision Transformer pipeline (DINO-pretrained ViT-B/16) benchmarked against a re-implemented REMEDIS ResNet-50 baseline, adding a hybrid consensus mode that surfaces cross-model agreement and disagreement.
- Engineered a data-orchestration pipeline over PadChest (160K+ X-rays, 67K+ patients), pairing frontal-lateral studies by StudyID and applying Weighted BCE loss + OneCycleLR to counter extreme class imbalance.
- Deployed a clinical LLM platform (LLaMA 3.3 70B + Whisper Large V3 via Groq) automating transcription, clinical note generation, and risk stratification.

RAG-Powered Medical Report Interpreter | *FastAPI, Pinecone, Next.js, MongoDB* Feb 2026 – Apr 2026

- Built a full-stack AI SaaS product interpreting medical lab reports through a 6-stage LLM pipeline (OCR → retrieval → analysis → trend detection → infographics) for clinician- and patient-friendly insights.
- Designed a scalable FastAPI backend with JWT auth, async MongoDB, Pinecone vector search, and rate limiting to support secure, low-latency AI inference.
- Deployed a Next.js frontend (Vercel) and FastAPI backend (Hugging Face Spaces), enabling cloud-hosted uploads, subscription management, and real-time analysis.

Hybrid Movie Recommendation System | *PyTorch, Sentence-BERT, NCF* Nov 2024 – Jan 2025

- Improved recommendation quality 15% over TF-IDF + cosine baselines by building a hybrid recommender combining all-MiniLM-L6-v2 embeddings with Neural Collaborative Filtering.
- Achieved RMSE ≈ 1.08 , Precision 0.80, and MAP 0.61 on a 2.5M+ interaction dataset, outperforming ALS and Transformer baselines, and resolved cold-start via content-based fallback. (*Manuscript under review.*)

Zero Trust Security in AI-Driven Healthcare Systems | *Python, Security* Apr 2025 – May 2025

- Reduced simulated DoS attack success rates by 95% by establishing a verifiable model-deployment trust chain (SHA-256 integrity hashing, Zero Trust-aligned traffic filtering) around AI backends for disease prediction achieving 90%+ accuracy.

RESEARCH

HAVIT: Attentive Probing of a Frozen Chest X-ray Foundation Model (*Under Review*) 2026

- Lead Researcher & Research Direction Lead on multi-view PadChest classification — submitted to FIT 2026 (23rd Intl. Conference on Frontiers of Information Technology, IEEE Technical Co-Sponsor, COMSATS University Islamabad); directed the full pipeline from problem formulation through statistical analysis and manuscript preparation.
- Achieved a 6.24-point improvement in mean AUC over a matched frozen-feature baseline via a parameter-efficient attentive-probing framework, with statistically significant gains validated using paired seed-level and correlated ROC-AUC testing; a systematic ablation surfaced a negative-result finding that explicit cross-view fusion did not improve performance in this frozen, low-data regime.

FedGraph-LoRA: Hierarchical Preference Graphs and Federated LoRA Adaptation for Personalized LLM Agents

Submitted to FLLM 2026 — (Under Review) 2026

- Baig, O., Zafar, A., Ali, H., & Ali, A. — Achieved 88%+ preference alignment and F1 ≈ 0.93 over vector-only baselines by designing hierarchical graph-based preference modeling for **LLM agents**.
- Reached 86% adapter-selection accuracy by implementing dual LoRA adaptation with dynamic local/global/blended routing.
- Delivered 11% higher coverage at 77–78% end-to-end accuracy while preserving user privacy by integrating federated continual learning into the agent's adaptation loop.

HONORS & AWARDS

- **MITACS Globalink Research Internship (GRI) Award — 2026:** Selected competitively among top-ranked international students for a 12-week funded research internship at the University of Winnipeg, Canada (Prof. Bradley Rey), integrating an LLM into a next-generation smartwatch health application.
- **1st Place, Final Year Project Exhibition — UET Lahore:** Ranked #1 out of ~55 industry- and research-grade projects with AiroDx.
- **1st Place, Generative AI Internship Project — Xavor Corporation:** Top team out of 10+ competing teams.

LEADERSHIP

- **Lead Organizer, Cipher Craft 2024:** Coordinated protocol, graphics, and event operations for Pakistan's premier free hackathon, serving 300+ participants as part of the core leadership team.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, C#, C++, SQL

AI / LLM Engineering: LLM Agents, RAG, Fine-tuning, LoRA/PEFT, Prompt Engineering, Vector Databases (Pinecone), LangChain, LangGraph, Groq, Whisper

ML/AI Frameworks: PyTorch, TensorFlow, Hugging Face Transformers, Scikit-learn, OpenCV, timm

Core Concepts: Deep Learning, Generative AI, Computer Vision (ViT), Federated Learning, NLP, Recommendation Systems, Explainable AI

Backend & Deployment: FastAPI, Flask, MongoDB, Docker, MLOps, REST APIs, Vercel, Hugging Face Spaces, React/Next.js

Certifications: ML Specialization (DeepLearning.AI & Stanford) · Oracle OCI 2025 Data Science Professional · Oracle OCI 2024 Generative AI Professional