Zaraar Malik

Islamabad,Pakistan

+923319359430, zaraarmalik5303@gmail.com, www.linkedin.com/in/zaraar-malik-811675246

Education

BS Artificial Intelligence **2021-2025**

NUCES, Islamabad

Experience

AI Engineer - Apexbeat.ai (United Kingdom) Sept 2025 - Current

- Architected and deployed a scalable OCR pipeline for medical textbooks and complex PDFs, processing **50,000+ pages** with **96%+ extraction accuracy** and reducing manual data processing effort by 70% through automated layout parsing, table reconstruction, and validation layers.
- Designed and productionized a Retrieval-Augmented Generation (RAG) system using **locally hosted Small Language Model**, currently serving **2,000+** users with **99% uptime** and sub-1.2s average latency, while reducing hallucination rates by 28% and improving top-3 retrieval accuracy to **89%**.
- Conducted controlled benchmarking of SLM-based pipelines against large proprietary models from Google and Amazon Web Services, achieving 60% lower cost per 1,000 queries and significantly improved cost-performance trade-offs for production deployment.
- Built and optimized end-to-end AI infrastructure on AWS and cloud GPU platforms including RunPod and Vast.ai, achieving 3.2× faster inference throughput, **42% lower VRAM** consumption via FP16/INT8 quantization, and **38% reduction** in infrastructure costs through autoscaling and workload optimization.
- Engineered a **layout-agnostic** figure extraction pipeline for heterogeneous medical PDFs spanning 50,000 pages, combining a fine-tuned **YOLOv10l document layout model** with custom **mask dilation strategies** and dual-path text detection (scanned and digital PDFs); extracted **2,800+** labeled medical figures with **91% caption-match accuracy**, reducing manual curation effort by **65%** across a structured 50,000-entry processing manifest.

AI Intern - Systems Limited (Pakistan) June 2025 - Aug 2025

- Developed and deployed **real-time human action segmentation** pipeline using temporal modeling and frame-level feature extraction, achieving 92% frame-level accuracy and maintaining sub-40ms inference latency for edge-compatible deployments.
- Built end-to-end object detection and multi-object tracking systems using YOLO-based detectors integrated with Deep SORT and ByteTrack, improving **MOTA to 81%** and reducing ID-switches by 35% through Kalman filter–based motion prediction under occlusion and motion blur.
- Designed comprehensive evaluation and optimization frameworks leveraging mAP, IoU, MOTA, MOTP, precision-recall curves, and latency benchmarks, increasing detection accuracy by 14% and improving tracking robustness by **22% through structured ablation studies**.
- Conducted technical feasibility and scalability analyses for production AI deployments, reducing projected infrastructure costs by 30% while ensuring performance stability under high-throughput, real-time constraints.

AI Intern - AIO Silicon Valley Startup (Pakistan) June 2024 - Aug 2024

- Engineered scalable data pipelines handling both **10M+ large-scale records** and low-resource datasets (<5k samples), implementing robust cleaning, structuring, and reproducible training workflows, leveraged unsupervised and semi-supervised techniques (domain-adaptive pretraining, pseudo-labeling) to improve **low-data performance by 22%** while reducing annotation dependency.
- Fine-tuned a domain-specific Google BERT-based model for entity extraction from unstructured text, achieving **91% F1-score** and improving **precision-recall performance by 18%**, while optimizing training to reduce convergence time by 30% across varying data scales.

AI Research Assistant - FAST School of Management Sept 2024 - Mar 2025

- Architected and evaluated ML pipelines on financial datasets for churn prediction, credit risk stratification, and **investment pattern recognition**, employing feature engineering, imbalance mitigation, and **cross-validation protocols** across temporal and demographic splits.
- Engineered a production-grade **Retrieval-Augmented Generation** (RAG) system for personalized investment advisory, implementing hybrid retrieval over heterogeneous financial corpora, combining **dense vector embeddings** with structured data indexing to deliver region-aware, context-grounded recommendations across Pakistani socioeconomic segments.

Technical Projects

- **Snap Shop GenAI Fashion Synthesizer:** Developed a web-based virtual try-on platform leveraging **LoRA - finetuned latent diffusion models** for photo-realistic garment synthesis, enabling high-fidelity fashion visualization for e-commerce applications.
- **Multi-Agent Medical Reimbursement Assistant**: Architected a **multi-agent LLM** pipeline automating end-to-end medical receipt reimbursement workflows, incorporating inter-agent collaboration, structured validation logic, and fault-tolerant error handling to minimize manual overhead.
- **PhishNet Malicious URL Classifier:** Engineered a multi-class **URL threat detection** system using XGBoost with handcrafted lexical, host-based, and structural features to classify phishing, malware, spam, and defacement URLs with high discriminative precision.
- **TinyLLM RAG Chatbot:** Experimented with lightweight **LLM architectures** for retrieval-augmented generation, systematically **optimizing chunking strategies** and benchmarking accuracy-efficiency trade-offs to minimize deployment cost without degrading response fidelity.
- **Procedural Game Level Generation:** Trained **DCGAN** and **WGAN** architectures on Super Mario Bros level corpora to synthesize structurally coherent, playable game maps, supporting automated level design and procedural content generation workflows.

Technical Skills

- Programming Languages: Python, JavaScript, HTML, CSS, OpenMP
- Frameworks & Deployment: PyTorch, Scikit-Learn, Keras, FastAPI, Flask, Docker, AWS
- DataBases: QDRANT, Pinnecone, MongoDB, SQL

Certifications

- Deep Learning with Pytorch - IBM
- AI for Medical Diagnosis - Coursera
- Evaluating and Debugging Generative AI - DeepLearning.AI

Honors and Awards

- Data Quest, Nascon 2023: Winner of Pakistan’s largest AI Hackathon.
- AIO Hackathon, FAST 2023: 2nd Place in AI-focused hackathon by AIO.
- Dean’s List: Recognized 4 times for outstanding academic achievement
- Infyma Hackathon, FAST 02024: Runner up AI Hackathon by Infyma.