

Muhammad Zeeshan

AI Engineer

📞 03256257787 ✉ muhammadzeshan.covers@gmail.com 📍 Lahore, Pakistan
🌐 linkedin.com/in/muhammad-zeshan-6b66a7350 🐙 github.com/AIMindCrafter 📁 Portfolio

👤 Professional Summary

Result-oriented AI Engineer specializing in Generative AI, Large Language Models, and scalable machine learning solutions. Proficient in LLM fine-tuning QLoRA, Unsloth, building RAG architectures LangChain, vector databases, and deploying high-performance APIs FastAPI, Docker. Experienced in deep learning and computer vision model optimization, with a track record of driving system efficiency, model accuracy, and operational automation.

🏢 Work Experience

Algotix AI

Software Engineer Intern

Lahore, Punjab, Pakistan

Jan 2026 – Mar 2026

- Architected and optimized backend services using Python and FastAPI, reducing service response latency by 20% and enhancing endpoint throughput.
- Collaborated with cross-functional engineering teams to design AI deployments.
- Built robust validation, automated error handling, and structured logging workflows, decreasing system downtime and improving debug observability.

Deep Embed Lab

Machine Learning Engineer

Lahore, Pakistan

Dec 2024 – August 2025

- Designed and optimized predictive ML/DL models using Python, PyTorch, and TensorFlow, improving accuracy by 15% across core business metrics.
- Automated daily data ingestion, processing, and ETL pipelines using Linux Shell Scripting and Cron jobs, reducing manual operational overhead by 30%.
- Performed rigorous exploratory data analysis on multi-million row datasets, visualizing insights with Matplotlib and Seaborn to guide executive product decision-making.

🔧 Key Projects

MRI Brain Tumor Detection & Classification | *EfficientNetB2, PyTorch, Grad-CAM, Gradio* | 🐙 GitHub

- Trained an EfficientNetB2 classifier on 7,000+ MRI scans to classify tumor categories, achieving 97%+ validation accuracy.
- Integrated Grad-CAM explainability overlays for visual clinical interpretation and deployed the model on HuggingFace Spaces using Gradio.

AI Voice Assistant | *LiveKit, Deepgram, Cartesia, Qdrant, Python* | 🐙 GitHub

- Architected a real-time voice agent using LiveKit for low-latency streaming, Deepgram for ASR, and Cartesia for TTS synthesis.
- Integrated Qdrant vector search to enable context-aware semantic retrieval, optimizing the audio pipeline for sub-second response latency.

Multi-Agent Research Assistant | *LangGraph, LangChain, Tavily API, Gemini, Python* | 🐙 GitHub

- Built a cooperative multi-agent system using LangGraph for hierarchical search planning, information retrieval, drafting, and review.
- Designed state persistence to preserve context throughout deep-dive iterations, generating comprehensive reports with automated source citations.

Multi-PDF RAG Assistant | *LangChain, ChromaDB, Gemini API, Streamlit* | 🐙 GitHub

- Built an end-to-end multi-document RAG pipeline using LangChain, ChromaDB vector store, and Gemini API with source citation.
- Implemented semantic search with cosine similarity reranking and deployed a Streamlit UI supporting session history.

QLoRA Fine-tuned Multimodal LLM (Qwen 3.5 4B Vision) | *Qwen 3.5 / Qwen 2 VL, Unsloth, PEFT, HuggingFace* | 🐙 GitHub

- Fine-tuned Qwen 3.5 (4B) Vision on multimodal domain-specific datasets using QLoRA via Unsloth, cutting VRAM usage by 70% to train on free Kaggle T4 GPUs.
- Merged adapter weights and published the model to HuggingFace Hub with an interactive live Spaces demo for real-time evaluation.

⚙️ Technical Skills

Generative AI & LLMs: LLM Fine-tuning (QLoRA, PEFT), Unsloth, LangChain, Retrieval-Augmented Generation (RAG), Agentic AI, Prompt Engineering

Machine Learning & CV: PyTorch, TensorFlow, Scikit-learn, YOLOv8, EfficientNet, Computer Vision, NLP

Vector DBs & Databases: Qdrant, ChromaDB, PostgreSQL, SQLite

Tools & Backend Frameworks: FastAPI, Docker, Python, Git, GitHub, Linux CLI, Bash Scripting, Pandas, NumPy

Islamia University of Bahawalpur

Bachelor of Science in Computer Science

Bahawalpur, Pakistan

Jun 2022 – Jun 2026

Certifications

Generative AI & LLMs: Oracle Generative AI Professional, IBM Generative AI Engineering (Watsonx), Google Gemini Certified Student

Machine Learning & NLP: Stanford Machine Learning Specialization, DeepLearning.AI NLP Specialization, IBM AI Engineering (PyTorch/TensorFlow)

Professional Tracks: DataCamp AI Engineering Track, DataCamp AI Engineer Associate, Microsoft AI & ML Engineering