

Muhammad Huzaifa Zamir

Senior Generative AI Engineer & Technical lead

Phone: +923365258828 Email: huzafaakamhz@gmail.com

LinkedIn: <https://www.linkedin.com/in/huzafa-zamir-siddiqui/>

Website: <https://portfolio-jyv.pages.dev/>

SUMMARY

4+ years of experience designing and delivering enterprise AI solutions in regulated banking environments. Expertise spans LLMs, RAG, agentic AI, multimodal document intelligence, and secure AI infrastructure across on-premises and cloud platforms. Leads cross-functional teams from strategy through production, driving solution architecture, enterprise AI adoption, and governance aligned with ISO/IEC 42001.

CORE STRENGTHS

Backend & Software Engineering: Python | FastAPI | REST APIs | Docker

Generative AI: LLMs | RAG | Agentic AI | LangChain | LangGraph | MCP

AI Infrastructure & LLMops: vLLM | Nvidia Triton Inference Server | LMStudio | Ollama | Ray Serve

Data Science & Analytics: PyTorch | Pandas | Power BI | Tableau

Cloud Computing: Azure | AWS | Oracle | RunPod | Cloudflare | OpenAI | Anthropic Claude | Google Gemini

Data and Development Tools: Postgres | Qdrant | Snowflake | Microsoft SQL Server | MongoDB | Neo4j

Leadership: Technical Leadership | Stakeholder & Vendor Management | Solution Architecture | Mentoring | Training

PROFESSIONAL EXPERIENCE

Senior AI Engineer | [Meezan Bank](#) | Karachi

September 2025 - Present

- Established & led a cross-functional AI delivery team (1 Senior AI Engineer, 3 Junior AI Engineers, and 2 Project Managers), directing the architecture and delivery of in-house and vendor-led AI solutions across Operations, Compliance, Products, and Digital Channels.
- Architected and led development of a law-enforcement correspondence platform that reduced processing turnaround from one week to near real-time, enabled the reassignment of 12 staff to review activities, and helped eliminate an approximately 18,000-email backlog.
- Architected a multi-agent transaction-investigation platform for automated counterparty mapping across account transactions, nostro statements, and customer data; currently in staging
- Built an AI quality evaluation platform analyzing 20k to 40k multi-channel customer interactions per week using RAG-based LLM-as-a-judge evaluation to identify chatbot improvement opportunities.
- Led the architecture and vendor managed delivery of AI document-intelligence solutions for bank statement processing and account-opening automation, reducing statement processing time from ~25 minutes to 2 minutes and shortening turn around time from ~2 days to 30 minutes.
- Assisted vendor in architecting a multi-agent Virtual Relationship Manager for WhatsApp and web channels using agentic RAG, GraphQL, and speech technologies, including support for Urdu TTS model fine-tuning.
- Supported the bank's AI Management System and ISO/IEC 42001 alignment, and delivered Generative AI and Agentic AI training to executive groups of 10–30 participants and bank-wide sessions of up to ~300 attendees.

Data Scientist | [SOFTOO \(PRIVATE\) LIMITED](#) | Islamabad

July 2023 – September 2025

- Served as the organization's primary Generative AI engineer, partnering with backend, frontend, QA, and product teams to deliver enterprise RAG and agentic AI solutions from concept to deployment.
- Built [Yani](#) with LangGraph-based autonomous agents for document analysis, dashboarding, visualization, code execution, and intelligent assistance, with feedback-driven evaluation to improve retrieval and response quality.
- Developed AI-assisted regulatory workflows supporting control assessment and evidence processing aligned to ISO/IEC 27001/27002, ISO 9001, GDPR, PDPL, AAOIFI, and State Bank policies.
- Applied multimodal AI, OCR, and NLP to extract and transform information from unstructured documents, and automated data workflows using Microsoft Fabric, Power Automate, and AWS Glue; evaluated DataRobot and Dataiku for accelerated experimentation.
- Built FastAPI-based backend services and APIs supporting enterprise-grade AI applications across cloud and on-premises environments.

Earlier Experience

“Technical Consultant” at [Errands Services](#) (Part-time), Islamabad, from November 2024 – March 2025

“Data Scientist” at y/TECH, Lahore from March 2023 - July 2023

“Junior Data Analyst” at [Ihdinaa](#) (Part-time), Karachi from January 2021 – March 2022

“Middleware Engineering Trainee” at [Aion Digital](#) (Remote), Bahrain from July 2021 – September 2021

CERTIFICATIONS

- 2024: [Oracle Cloud Infrastructure 2024 Generative AI Certified Professional](#)
- 2024: [Microsoft Azure Fundamentals](#)
- 2023: [Snowflake Hands on Essentials](#)
- 2023: [Google Data Analytics Specialization](#)

Projects

- [MindMesh](#) - Local-first RAG knowledge assistant supporting private document ingestion, semantic retrieval, and local LLM inference.
- [onlocalai](#) - Local-first AI application marketplace for discovering and running private AI tools.
- [Shelfdom](#) - Enterprise document-intelligence and governance platform for multimodal extraction, review workflows, and controlled knowledge access.
- [Kinderview](#) - An augmented reality, AI assisted learning experience

Education

- **BS in Computer Science - 2022**
3.18 CGPA
Iqra University