MUHAMMAD FAZEEL YAQOOB

AI/ML ENGINEER | PRODUCTION LLM, RAG & MULTI-AGENT SYSTEMS

Kuala Lumpur, Malaysia | +60 19-8792523 | 6fazeel18@gmail.com | linkedin.com/in/fazeel-yaqoob

PROFESSIONAL SUMMARY

AI/ML Engineer with 5+ years of hands-on industry experience building production AI systems across LLM agents, RAG platforms, speech AI, document intelligence, analytics automation, and real-time computer vision. Currently pursuing a Master's by Research in Artificial Intelligence focused on privacy-preserving ML and federated learning. Strong at turning AI prototypes into governed systems with model routing, cost controls, auditability, role-based access, and human review. Builds end-to-end AI applications from backend architecture and model orchestration to APIs, data workflows, and user-facing product delivery.

CORE COMPETENCIES

Cloud Platforms & AI Services	AWS Bedrock, Lex, AWS AgentCore, EC2, Rasa, AWS Transcribe, AWS Connect, Azure AI Foundry, Azure OpenAI, Azure Speech, Google Vertex AI, Azure Document Intelligence, MCPs, Sagemaker, MLFlow
LLM Engineering & Agentic AI	RAG systems with self hosted LLM, prompt engineering, multi-agent orchestration, tool calling, structured outputs, model routing, LangChain, LangGraph, Strands, Embedding models
AI & Machine Learning	TensorFlow, Scikit-learn, Transformers, BERT, YOLO,PyTorch, OpenCV, model evaluation
Computer Vision & Multimodal AI	YOLO, OpenCV, OCR, MediaPipe, vision-language models, speech-to-text, transcript analysis, TTS
Programming & Development	Python, FastAPI, Flask, Next.js, Docker, Git, REST APIs, JWT auth
Databases & Data Engineering	Vector databases, PGVector, PostgreSQL, MySQL/RDS, Redis, DuckDB, FAISS, Chroma
AI Developer & Productivity Tools	Hermes , Amazon Quick, OpenClaw, Claude Code, GitHub Copilot, Microsoft Copilot Studio , n8n, Azure Logic Apps, Power Automate

PROFESSIONAL EXPERIENCE

Premier NX — AI Engineer **Nov 2023 – Present**

Kuala Lumpur, Malaysia / Hybrid

- Architected production AI systems across multi-agent analytics, QA evaluation, RAG workspaces, document intelligence, and voice/speech workflows.
- Built a self-service analytics platform that reduced reporting cycles from about 2 hours to under 5 minutes using deterministic analytics plus LLM agents.
- Developed a multi-client QA evaluation system that transcribes calls, scores configurable rubrics, and reduced manual QA workload by 80%.
- Designed secure RAG and document-intelligence workflows with workspace isolation, model routing, token/cost tracking, audit logs, and human review.
- Integrated AWS Bedrock, AWS Transcribe, Azure OpenAI, Azure AI Foundry, Deepgram, LangChain, FastAPI/Flask, and Docker-based deployments.

VdoTok — Data Scientist **Apr 2022 – Nov 2023**

Lahore, Pakistan

- Implemented NLP pipelines and conversational AI using Transformers, LangChain, and custom BERT-based models.
- Developed real-time speech recognition and sentiment analysis systems for communication insights.
- Optimized ML inference pipelines in Dockerized environments for scalable production deployment.
- Worked on data preprocessing, model retraining, and feature engineering to improve accuracy and performance.

Norgic AB — Software Engineer (Data Science) **Sep 2021 – May 2022**

Lahore, Pakistan

- Contributed to AI-powered analytics and computer vision pipelines for automation and image classification tasks.
- Built data ingestion and preprocessing workflows supporting the end-to-end ML lifecycle.
- Collaborated with R&D teams to train and deploy models using TensorFlow and PyTorch.

TECHNICAL PROJECTS

Data Project Analyzer — Multi-Agent Self-Service Analytics Platform

- Built a multi-agent analytics platform that ingests CSV/XLSX/JSON files, runs an 18-agent pipeline, and delivers interactive dashboards with grounded chat and code execution.
- Implemented task-aware LLM routing through AWS Bedrock, keeping deterministic calculations in code while reserving larger models for synthesis and reasoning.
- Enabled parallel multi-file processing with ThreadPoolExecutor, real-time SSE streaming, and an in-chat code interpreter for on-the-fly data queries and Plotly visualizations.

  Tech: Python, Flask, AWS Bedrock, LangGraph, AWS AgentCore, Polars, DuckDB, Plotly, Redis, SSE

Premier Nexus — Enterprise Document Intelligence & Conversational AI Platform

- Architected a multi-tenant RAG platform with document ingestion (PDF, DOCX, CSV), OCR extraction, vector search via PGVector, workspace-isolated chat, and streaming responses.
- Integrated multiple model providers with per-token cost tracking, model-aware routing, retries, API controls, and role-based workspace access.
- Built enterprise features including API key management, rate limiting, speech-to-text/TTS support, usage logs, and 54 active backend routes.

  Tech: Python, FastAPI, LangChain, AWS Bedrock, Azure OpenAI, PGVector, Azure Blob Storage, Docker, ElevenLabs

PremierNX QA Evaluation System — AI-Powered Call Center Scoring Platform

- Built a multi-client QA evaluation platform that transcribes call recordings, scores agents against customizable criteria, and generates evidence-backed scorecards, reducing QA workload by 80%.
- Integrated 3 pluggable STT providers (Deepgram, AWS Transcribe, EdenAI) with automated PII redaction, and AWS Bedrock LLMs (Claude, Llama, Nova) for evidence-based scoring with per-token cost tracking.
- Designed multi-client database architecture supporting unlimited clients/forms, role-based access, audit logging, and enterprise JWT authentication with PremierNX portal.

  Tech: Python, Flask, AWS Bedrock, Deepgram, AWS Transcribe, MySQL/RDS, Docker, LangChain, S3

Premier Scan — Intelligent Logistics Invoice Processing System

- Built a document intelligence system that combines GPT-4o Vision with Tesseract OCR for cross-verified extraction of shipping invoices, converting multi-page PDFs into CASS-standard 45-column CSV exports — improving data entry efficiency by 80%.
- Implemented fuzzy charge-code matching (85% threshold), multi-currency normalization, 250+ country code lookups, and per-document cost tracking.

  Tech: Python, Flask, Azure OpenAI (GPT-4o Vision), Tesseract OCR, SQLite, Docker

EDUCATION

Universiti Kuala Lumpur (UniKL MIIT) Sept 2025 – Present

Master of Information Technology (by Research) — Artificial Intelligence Focus

Kuala Lumpur, Malaysia

- Research Focus: Privacy-Preserving ML and Federated Learning
- Thesis focus: Deployment-Ready ML Systems for Enterprise Applications

The University of Lahore Graduated: 2020

Bachelor of Science in Computer Science

Lahore, Pakistan

- Final Year Project: Wheat Disease Detection using YOLO with Mobile App Integration
- Relevant Coursework: Machine Learning, Deep Learning, Natural Language Processing

CERTIFICATIONS

- AWS Training & Certification — Build Strands Agents with SageMaker AI Models and MLflow (2026)

- AWS Training & Certification — Optimizing Foundation Models; Essentials of Prompt Engineering; Fundamentals of ML and AI
- Anthropic — Claude Code in Action (2026)
- Google Cloud — Prompt Design in Vertex AI; Explore Generative AI with the Vertex AI Gemini API; Using the Google Cloud Speech API
- DeepLearning.AI / Stanford Online — Machine Learning Specialization
- LinkedIn Learning — Advanced NLP with Python for Machine Learning; Advanced Prompt Engineering Techniques
- Coursera / DeepLearning.AI — Supervised Learning, Advanced Learning Algorithms, Unsupervised Learning & RL