Muhammad Ahmad

AI/ML Engineer · LLMs & RAG · Multi-Agent Orchestration

+92 309 1148223
avcton.github.io
avcton@gmail.com
linkedin.com/in/avcton
Lahore, Pakistan · Open to Remote

SUMMARY

AI/ML Engineer with production experience building multi-agent orchestration systems, RAG pipelines, and LLM-powered automation at Dubizzle Labs, the tech arm behind Bayut and OLX in MENA. Architected LangGraph-based agentic infrastructure on Kubernetes/EKS, led OpenAI Responses API migration across a multi-tenant ML platform (~200 model configs), and deployed AI agents achieving **6.2% conversion rate** and **400% outperformance vs telesales**. Gold Medalist, Summa Cum Laude, FAST NUCES; first in the BS Data Science batch.

EXPERIENCE

Machine Learning Engineer

April 2025 – Present

Dubizzle Labs — Lahore, Pakistan (On-site)

- Built Nexus: FastAPI + LangGraph + MCP Server gateway providing unified customer intelligence over MindsDB, with a ReAct agent loop and dual REST + MCP interface for cross-platform AI memory across Bayut and Zameen
- Led R&D evaluation of multi-agent orchestration frameworks (GoClaw, LangGraph, CrewAI, and 3 others) for a 100–500 consultant AI agent fleet; presented live demo to CEO in a 30+ person boardroom; **CEO approved GoClaw as company-wide go-forward direction** for the Agentic Sales Platform
- Migrated Ingress (Dubizzle's multi-tenant ML platform, Rails 7) from OpenAI Chat Completions to Responses API across **~200 model configs**, enabling GPT-5 family support; identified and fixed a critical production routing bug during migration
- Architected n8n as the central orchestration layer for two production AI products; designed PostgreSQL/Redis queue mode, inverted GitHub Actions CI/CD pipeline, and shared Postgres Chat Memory with cross-channel session key convention; deployed to Kubernetes/EKS
- Trained custom multi-head Longformer BERT (13 output heads) for LLM hallucination detection in call transcriptions; also built LangGraph agentic pipeline as an alternative; selected BERT for production scalability and cost efficiency
- Built and led AI calling agents across **6 production campaigns** (peak: 22,350 calls/campaign), achieving **6.2% conversion rate** and **400% outperformance vs telesales**; drove **40% improvement in call quality**
- Engineered Zara, Bayut KSA's Arabic-dialect AI calling agent for live real estate expo campaigns in Riyadh and Jeddah; built knowledge base, geopolitical resilience layer, and venue navigation guide

AI Research Engineer (Contract)

October 2024 – June 2025

Nextbridge Limited — Lahore, Pakistan

- Led end-to-end development of Interview Sensei: AI-powered digital human mock interview platform with 3D talking avatar, real-time WebRTC video/audio, LangChain-driven dynamic question generation, and post-session performance analytics
- Solo-deployed on RTX 4090 via NGINX reverse proxy with GPU session rate-limiting; won **1st place at FAST NUCES FYP Competition** (250+ projects) and **1st place SOFTEC 2025** (Pakistan-wide)

Full Stack Developer Intern

January 2024 – August 2024

Axcel Sea Shipping Co. LLC — Dubai, UAE (Remote)

- Built Flutter-based service app; integrated Dockerized Node.js/Express backend with MongoDB; **improved DB query execution by 30%** using Mongoose ORM; contributed to 20+ API endpoints across 10+ sprint cycles

PROJECTS

Interview Sensei: Your Goto Interview Coach
November 2024 – April 2025

Final Year Project — Sponsor: Nextbridge Ltd.

- AI-powered mock interview platform: users upload a resume, provide a JD, and conduct a live session with a 3D avatar; delivers post-session analytics on confidence, speaking rate, engagement, and answer accuracy
- Stack: FastAPI, LangChain, Llama 3.3 70B, WebRTC, Next.js, OpenCV, MongoDB, NGINX, RTX 4090

EDUCATION

B.S. Data Science
September 2021 – June 2025

FAST NUCES, Lahore — CGPA 3.90/4.00, Summa Cum Laude, Gold Medalist, 1st in batch

SKILLS

AI / ML	Python, LLMs, LangGraph, LangChain, RAG, Multi-Agent Systems, FastAPI, Prompt Engineering, Fine Tuning, NLP, Generative AI, Hallucination Detection
Orchestration	n8n, LangGraph, MCP Server, RabbitMQ, Celery, Sidekiq, GitHub Actions CI/CD
Infrastructure	AWS (EKS), Kubernetes, Docker, NGINX, PostgreSQL, Redis, MongoDB, MindsDB, Elastic Stack
Voice AI	ElevenLabs Conversational AI, SIP Trunking, Telnyx, SIP/RTP, DID/VoIP, TTS/STT, Faster Whisper
Languages	Python, Ruby, TypeScript, JavaScript, Dart, Go, SQL
Frameworks	Ruby on Rails, Next.js, Flutter, React, Node.js, Three.js

LANGUAGES

English — Professional **Urdu** — Native **Punjabi** — Native