Huzaifa Tahir

AI Engineer & AI Researcher — Generative AI · LLM · RAG · Multi-Agent Systems

Lahore, Pakistan

+92 321 8504372 | huzaifatahir7524@gmail.com | huzaifatahir.com

linkedin.com/in/huzaifa-tahir-ai | github.com/Huzaifa-X | medium.com/@huzaifatahir7524

Summary

AI Engineer and applied AI Researcher with 3+ years of production experience building **agentic AI systems**, **multi-agent orchestration**, **RAG pipelines**, and **LLM-powered automation** for legal, finance, sales, and HR domains. Delivered a **top-ranked Deep Research Sales Intelligence Agent** for enterprise clients and cut payroll processing time by **90%** through LLM-driven automation. Proficient in Python, CrewAI, LangGraph, LangChain, LlamaIndex, and **MCP (Model Context Protocol) Servers**, with end-to-end ownership from architecture and backend development to AWS deployment. Published author on LLMs, NLP, and retrieval systems.

Technical Skills

Programming Languages: Python, SQL, R
Generative AI & LLMs: OpenAI (GPT-4, GPT-4o), Anthropic Claude, Groq, Deepgram, Upstage.ai, Tavily, LLaMA 3, DeepSeek, Qwen 2.5, Kimi
Agentic AI & Orchestration: CrewAI, LangGraph, LangChain, LlamaIndex, LangSmith, MCP Servers (Model Context Protocol), AI Agents, Multi-Agent Systems, Prompt Engineering
RAG & Vector Databases: Retrieval-Augmented Generation, ChromaDB, Pinecone, Pgvector, FAISS, Hybrid Search, Re-ranking
ML / DL / NLP: PyTorch, TensorFlow, Hugging Face, scikit-learn, Deep Learning, Natural Language Processing, Computer Vision, Fine-tuning
Backend & APIs: FastAPI, Django, Django REST Framework (DRF), REST APIs, Microservices
Cloud & Databases: AWS (EC2, S3), Docker, PostgreSQL, SQLite, MySQL, Snowflake
Tools & Automation: Git, GitHub, Postman, VS Code, Jupyter Notebook, Playwright, ComfyUI, Streamlit

Work Experience

Musketeers Tech (Private) Limited — **Apr. 2025 – Present**
AI Engineer — *Onsite — Lahore, Pakistan*

- **Deep Research Sales Intelligence Agent:** Architected and delivered a full-stack agentic AI system using **CrewAI** and **FastAPI**, enabling autonomous multi-step research to identify and qualify B2B sales leads — **top-ranked result** among client deliverables; significantly reduced manual prospecting effort.
- Engineered an automated **payroll and salary-slip generation system**, cutting manual processing time by **90%** and eliminating calculation errors across the entire workforce.
- Developed custom **MCP (Model Context Protocol) Servers** integrated with OpenAI, enabling modular AI workflow orchestration and reducing agent integration overhead across projects.
- Deployed and optimized **ComfyUI** pipelines for AI image generation, automating creative production workflows and reducing turnaround time for visual assets.
- Built automated **Playwright** web scrapers with robust error handling, extracting structured data at scale to feed downstream AI processing pipelines.
- Designed **multi-chatbot agent systems** serving diverse business use cases, delivering plug-and-play conversational AI across multiple departments.

Fabulous Technology Solutions (Fabtechsol) — **Oct. 2024 – Apr. 2025**
AI Engineer — *Onsite — Lahore, Pakistan*

- **B-Master Multi-Agent Platform:** Led end-to-end integration of a ChatGPT-style **multi-agent platform** using **CrewAI**, **LangChain**, and **LangGraph**, delivering advanced analytics and intelligent automation across multiple agent stores and resource hubs.
- Fine-tuned **GPT-4 Vision** for car model and year detection from images, achieving **74% classification accuracy** and enabling downstream automated price estimation for vehicle inventory.
- Developed AI agents for tax-strategy recommendations using **LangChain SQL Agent**, OpenAI, and Tavily Search, streamlining financial advisory workflows and reducing analyst research time.

- Engineered utility-bill data extraction pipelines integrating **Upstage.ai** and GPT-4 with confidence scoring, improving data precision for downstream billing analytics.

Mexa Solutions **Sep. 2023 – Oct. 2024**

Generative AI Engineer *Onsite — Lahore, Pakistan*

- Built an **Interview Copilot** powered by OpenAI, **Deepgram** (real-time STT), and Groq LLM, delivering live conversational coaching with sub-second response latency.
- Implemented advanced **RAG pipelines** using LangChain and LlamaIndex for cybersecurity document analysis and academic research tools, improving retrieval relevance for domain-specific queries.
- Designed **"Study Buddy,"** an AI-powered personalized learning assistant generating adaptive study plans and condensing academic content to improve student productivity.
- Built a **Finance Streamlit Chatbot** with **Snowflake** integration, enabling natural-language querying of financial data warehouses for real-time business insights.

FalconXoft **Dec. 2022 – Aug. 2023**

Associate Python Developer (Internship → Full Role) *Remote*

- Contributed to a **Flight Reservation Management System** integrating third-party Flight APIs — handled booking logic and backend API development in Python.
- Developed a **Django application** for real-time face-emotion recognition using CNNs, demonstrating early proficiency in computer vision and ML deployment.
- Built a **web scraper** to compute Best Seller Rank (BSR) for e-commerce products, delivering data-driven insights that informed merchandising decisions.

Projects

EDA LangChain Agent | *Python, LangChain, OpenAI, Streamlit, SQL* **GitHub** | **2024**

- Natural-language interface for Exploratory Data Analysis — users query datasets in plain English to receive statistical summaries, correlation heatmaps, and ML-ready insights via the **LangChain SQL Agent**.

AI Study Plan & Book Summarization | *Python, LangChain, ChromaDB, GPT-4, Groq* **GitHub** | **2024**

- Generates personalized study plans and condenses full-length books into structured summaries using chunked **ChromaDB** retrieval to overcome GPT-4's context-window constraints.

Arabic Speech Recognition — Whisper Fine-Tuning | *Hugging Face, OpenAI Whisper, PyTorch* **GitHub** | **2023**

- Fine-tuned OpenAI Whisper on a curated Arabic speech dataset using Hugging Face Transformers, achieving accurate Arabic ASR transcription benchmarked on held-out test audio.

YourStudyBuddy — School Chatbot Platform | *OpenAI, LangChain, Django, RAG, AWS EC2* **Live Site** | **2023**

- Deployed a **RAG-based chatbot** for an American high school, supporting PDF uploads, automated question generation, and staff–student communication — hosted on AWS EC2 for production availability.

Publications

LLM: Large Language Models — A Comprehensive Guide | *AI in Plain English* **2024**

Understanding Seq2Seq Models: Revolutionizing Language Processing | *AI in Plain English* **2024**

Unleashing the Potential of Language: Introducing PaLM | *AI in Plain English* **2024**

Decoding the Magic: NLP Tokenization and Text Normalization | *Medium* **2024**

Education

Superior University Lahore **2019 – 2023**

Bachelor of Science in Computer Science *Lahore, Pakistan*

- **Relevant Coursework:** Data Science, Machine Learning, Database Systems, Software Engineering, Artificial Intelligence.

Certifications

Machine Learning Specialization **2023**

Coursera — DeepLearning.AI

Natural Language Processing Specialization **2023**

Coursera — DeepLearning.AI