

TALHA ZAIN

Applied AI Engineer | Agentic AI, LLM Systems & RAG

Islamabad, Pakistan | +92 340 5555550 | talha.10.zain@gmail.com | [linkedin.com/in/talhazain10/](https://www.linkedin.com/in/talhazain10/)

PROFESSIONAL SUMMARY

Applied AI Engineer with 3+ years of experience designing and shipping production AI systems across agentic workflows, retrieval-augmented generation (RAG), multi-agent orchestration, generative media, semantic search, conversational AI, and machine learning. Hands-on with LangGraph-based agents, LLM evaluation and observability, vector retrieval, FastAPI services, asynchronous workflows, and containerized cloud deployments. Built systems spanning healthcare, advertising, multimodal content generation, voice AI, and data/ML platforms, with emphasis on reliability, measurable quality, latency, and production readiness.

CORE TECHNICAL SKILLS

Agentic & GenAI	LangGraph, LangChain, Agentic RAG, Multi-Agent Systems, Tool/Function Calling, Human-in-the-Loop, Prompt Engineering, Structured Outputs
LLM & Evaluation	OpenAI, Claude, Llama, Hugging Face, LangSmith, RAGAS, Embeddings, Retrieval Evaluation, LLM Observability
ML & Data	PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost/LightGBM, NLP, Computer Vision, Classification, Regression, Clustering, Forecasting, SQL
Backend & Retrieval	Python, FastAPI, Flask, REST APIs, Async Workflows, PostgreSQL, pgvector, Redis, MongoDB, Vector Databases, Geospatial Search
MLOps & Cloud	Docker, CI/CD, AWS, GCP, Azure, Model Deployment, Monitoring, API Integration, Batch/ETL Pipelines

PROFESSIONAL EXPERIENCE

BoolMind — Lahore

Jul 2025 – Present

ML Engineer — Applied / Generative AI

- Architected an LLM-powered semantic advertising recommendation engine that converts natural-language campaign requirements into structured geo-intent, validates locations semantically, and ranks OOH inventory using footfall (30%), traffic (30%), POI density (20%), and proximity (20%) across 290+ UAE categories, returning optimized recommendations in under 5 seconds.
- Designed production AI services with Python, FastAPI, vector retrieval, Redis/PostgreSQL/MongoDB patterns, Docker, and cloud deployment workflows; integrated external APIs with retry/backoff and idempotent processing for reliable AI and data pipelines.
- Built and evaluated agentic/RAG workflows using LangGraph, retrieval pipelines, tool-based execution, stateful routing, and LLM observability/evaluation practices including LangSmith and RAGAS.
- Applied ML and optimization techniques across supervised/unsupervised learning, recommendation, forecasting, and search problems, focusing on production quality rather than notebook-only experimentation.

IT Genics — Lahore

Jan 2024 – Jun 2025

Associate AI Engineer

- Built a no-code conversational AI platform enabling organizations to configure and launch custom domain assistants without writing code, reducing dependence on bespoke chatbot development.
- Developed LLM-powered conversational and voice workflows integrating Llama, NLP pipelines, Twilio, APIs, and business actions for support, booking, information retrieval, and automation use cases.
- Designed and deployed AI/ML solutions spanning predictive analytics, NLP, computer vision, and chatbot automation, packaging models behind production APIs and containerized services.
- Collaborated across product and engineering teams to translate business requirements into deployable AI workflows, data pipelines, validation logic, and measurable system behavior.

IT Solutions World Wide — Islamabad

Jul 2023 – Dec 2023

Junior AI Engineer

- Built NLP applications using GPT-3, Llama, BERT, and Falcon, applying transfer learning/fine-tuning and prompt-based techniques to domain-specific conversational and language-processing use cases.
- Worked with TensorFlow, PyTorch, neural networks, APIs, and end-to-end AI pipelines to move experiments into reusable services and support cross-functional delivery.

PRODUCTION AI PROJECTS

AI Compare Hub — Multi-Model Generative AI Platform | ai-compare-hub.com/

- Engineered a production multimodal generation platform spanning 30+ image models, 20+ video models, and 5+ audio models, with unified access to providers including Flux, Kling, Veo, Seedream/Seedance, Runway, Luma, Ideogram, Stable Diffusion, MiniMax, and OpenAI.
- Supported simultaneous comparison of up to 4 AI models from a single prompt plus multimodal workflows including text-to-image, image-to-image, text-to-video, image-to-video, video-to-video, extend/merge video, text-to-music, text-to-speech, and audio-video generation.
- Implemented asynchronous generation architecture with queue-based processing, webhook/polling patterns, retries, status tracking, cloud asset handling, generation history, and model-specific parameter preservation across staging and production.

MedicAI — Multi-Tenant Multi-Agent Clinical AI Platform

- Designed a LangGraph-based multi-agent architecture with specialized clinical agents for tasks such as ICD-10 coding, SOAP documentation, scheduling/booking, medical imaging, triage, and clinical Q&A, backed by Agentic RAG and tool-based workflows.
- Implemented production controls including tenant isolation/RBAC, Redis + PostgreSQL state/memory patterns, retrieval pipelines, structured routing, human-in-the-loop gating, schema validation, guardrails, and evaluation/observability using LangSmith and RAGAS.

MediaX — Semantic & Geospatial OOH Advertising Recommendation

- Built an LLM-driven search and ranking pipeline that parses campaign intent, maps locations and POI types semantically, performs geospatial retrieval, and scores advertising inventory across 290+ UAE categories using weighted business signals.
- Combined API/data enrichment, retry/backoff, caching, conflict resolution, idempotent upserts, and configurable ranking logic to produce optimized recommendations in under 5 seconds.

ECG Intelligence Platform — Clinical Data Unification & ML Readiness

- Unified heterogeneous PDF, image, CSV, and waveform inputs into a vendor-agnostic canonical schema with configurable field mapping, enabling downstream analytics and machine-learning training pipelines.
- Automated data normalization and mapping workflows that eliminated approximately 3–4 hours of recurring manual analyst effort per week.

EDUCATION & CERTIFICATIONS

National University of Computer & Emerging Sciences (FAST-NUCES), Islamabad

Bachelor of Computer Science | Aug 2019 – Jun 2023

Certifications:

Microsoft Azure Fundamentals (AZ-900)	IBM Machine Learning with Python
IBM Databases & SQL for Data Science	IBM Data Analysis with Python
IBM Data Science Methodology	Tools for Data Science