

TALHA ANWAR

AI / ML Engineer — Generative AI, LLM Systems & Full-Stack Delivery

AI/ML engineer with 6 years shipping production AI end to end — LLM/RAG applications, AI agents & agentic workflows, prompt engineering, fine-tuning, and LLM evaluation/observability — with full-stack ownership across Python (FastAPI/Django) and Node.js. Founding engineer who scaled a conversational-AI product to 30+ B2B clients and **€400K ARR**, and solo-built & operates **AcruxCore**, a live LLMops platform (a LangSmith / Langfuse competitor). Ships at startup speed pairing strong manual engineering with AI-augmented workflows (Claude, Cursor). Proven with PyTorch, Hugging Face, and vector databases in production. **Open to remote (worldwide) and relocation (UAE/Gulf, Europe).**

CONTACT

Email: aiengr.talha@gmail.com — Phone: +92-331-9709944

Web: talhaanwar.com — github.com/talhaanwarch

PROFESSIONAL EXPERIENCE

Founding & Lead AI / Full-Stack Engineer (Remote) — MayaTravel (Belgium) Jul 2023 – Present

- Founding engineer on a **conversational & agentic AI** product built on LLMs; scaled from first B2B client to **30+ enterprise implementations**, **0% churn**, and **€400K ARR**, working fully remote across time zones.
- **Own the product end to end as full-stack engineer:** build the **core agentic features and LLM agents**, the **database**, **client-facing dashboards**, the **prompt systems**, and the **data-preparation and feedback-handling / evaluation pipelines** that keep model quality improving in production.
- Architected an advanced **RAG pipeline** over multimodal data (text, images, structured): **intent detection** triggering specialized prompts, hybrid metadata search with fallbacks, and API-first / vector-store-first variants per client; built **hallucination-prevention (guardrails)** and abuse-detection with quantitative evaluation.
- Led migration of the Python AI service → **Node/NestJS**; deployed an **Airflow pipeline on Azure** for automated data extraction and vector-store updates.

NLP Consultant (Remote) — Soffos.ai (USA)

Oct 2022 – Mar 2023

- Engineered GPT-3 prompts and built **FastAPI** microservices for NLP tasks (classification, ambiguity detection) with spaCy/NLTK, integrated into the Soffos SDK.
- Built a document parser extracting structured data from complex layouts.

Founding Engineer / Data Scientist — EzShifa (Pvt) Ltd (Pakistan) Aug 2020 – Sep 2022

- Led AI for a telemedicine startup: remote vital-sign detection with **PyTorch**/TensorFlow from phone-camera video (± 5 MAE on heart rate).
- As founding engineer, set up the **AWS infrastructure** end to end — **EC2** (web app), **Lambda**, **S3** (video storage), and SageMaker with Docker, scaling to concurrent video consultations.

SELECTED AI PROJECTS

- **AcruxCore.com** — **LLMOps Platform** (live product, a [LangSmith](#) / [Langfuse](#) / **Prompt-Layer competitor**): solo-built and operate a production LLMOps control plane — prompt versioning with live aliases, an OpenAI-compatible **gateway** (routing/cost/caching), per-call **tracing**, versioned tools, and dataset-driven **evaluation** / **A-B testing**; ships **Python and Node.js SDKs**. I set up the **CI/CD**, database, **auth**, and **AWS SES**, and run the whole stack.
- **Autovidify.com**: independently architected and shipped a full-stack GenAI SaaS (multi-provider ASR + context-aware translation) — Python AI core, Node/Express API, React, Supabase auth, AWS S3/SES, Redis — to production in under 2 months.
- **AfterLib.com** — **Ad Quality Classification**: built and trained the classification system behind ad-quality scoring — **500K+ ads** across **100 categories** at **F1 82%**, with a data pipeline built for that scale and delivered production-ready.
- **Glendor.com** — **PHI Sanitizer**: Protected Health Information redaction combining **YOLO** detection with **NER** and statistical pattern recognition, reaching **80%** accuracy on sensitive-data detection.

TECHNICAL SKILLS

- **Generative AI & LLMs**: LLMs (OpenAI & open-source), RAG (specialized implementations), AI Agents & Agentic Frameworks, LangChain, LangGraph, Prompt Engineering, Fine-tuning (PEFT), LLM Evaluation, Guardrails, LLMOps
- **Backend & Full-Stack**: Python (expert), FastAPI, Django, Node.js, NestJS, Express, TypeScript, React, RESTful APIs, Docker, CI/CD
- **Deep Learning & NLP**: PyTorch, TensorFlow, Hugging Face, Transformers (BERT, GPT, LLaMA), spaCy, NLTK
- **Data & Cloud**: Vector Stores (Milvus, Qdrant, FAISS), PostgreSQL, Airflow, AWS (S3, SES, EC2, Lambda, SageMaker), Azure, GCP, Redis, Sentry
- **AI-Augmented Delivery**: Claude, Cursor — compressing multi-day tasks into hours at startup speed

EDUCATION & RESEARCH

- **MS Data Science**, Fast National University, Pakistan 2018–2021
- **23 peer-reviewed papers, 580 citations, h-index 12, i10-index 15** on ML/DL — [Google Scholar](#); multiple international ML competition wins.