

MUHAMMAD BILAL TAHA

Full Stack AI Engineer | RAG, LangGraph, FastAPI, Next.js

☎ +92-370-814-8038 ✉ m.bilal.25132@gmail.com 🌐 mbilaltaha.com 📄 in/bilal-taha25132 🌐 bilaltaha25132

Experience

EdgeFirm

Mar 2026 – Present

Full Stack AI Engineer

Karachi, Pakistan

- **DCT Digital Concierge** (Abu Dhabi Dept. of Culture & Tourism, with **Create**, Dubai): **LangGraph** ReAct agent on **GPT-4o** orchestrating **15+** tools (semantic search, geo-proximity, Maps routing, itinerary generation), streaming typed payloads over **SSE** so the UI renders venue cards and curated itineraries rather than plain text.
- Owned the retrieval layer end to end: multi-format ingestion (PDF, DOCX, PPTX, HTML) from **Azure Blob** chunked into **pgvector** on a scheduled re-index, keeping answers grounded in curated tourism content instead of model memory on a public-facing government product.
- Hardened it for public deployment with multi-layer **prompt injection defense**, content moderation, and **Azure OpenAI** fallback; traced every LLM call via **Langfuse** and **Arize Phoenix**, with conversation state persisted in **Postgres** for multi-turn continuity.
- Shipped the chat UI as a self-contained **Next.js** widget embedded into a legacy **Sitecore** property and fully decoupled from its CMS; converted Figma designs to production components via **Figma MCP servers**.

Associate Software Engineer

Nov 2024 – Mar 2026

- **Harbinger** (market intelligence, UK-based food ingredients firm): **FastAPI** backend tracking **35+** companies across 4+ data sources for the client's executive board, cutting competitor signal discovery time by **~80%**.
- **ReAct agentic RAG engine** (**LangGraph** + **GPT-4o**, 13 tools, **pgvector**) with dual **SSE** streaming for chat and live notifications, cutting analyst query resolution time by **~70%**; similarity dedup collapses near-duplicate stories before the LLM layer.
- Async-first ingestion via **Celery** (4 queues, 10+ cron jobs) that autonomously scrapes, detects, classifies, and alerts at **<5-minute** end-to-end latency with zero human intervention.
- **Medianet** (K-Electric, Pakistan's largest power utility): **NestJS** field operations backend processing **1,000+** cases daily across **300+** field officers with **9-level** role-based access and GPS tracking at **2,000 req/min**, fronted by an offline-first **Next.js** PWA using **IndexedDB** to eliminate data loss in low-connectivity field conditions.

Projects

Manifexa: AI Voice Ritual and Wellness Platform | *NestJS, OpenAI TTS, BullMQ, PostgreSQL*

- **Voice synthesis pipeline** on **OpenAI TTS** (gpt-4o-mini-tts) mapping **8 tones** across **9 voices** to per-request prosody instructions, with a live voice-preview endpoint and graceful degrade when synthesis fails.
- **NestJS** backend of **161 REST endpoints** over **30 Postgres** models; central **OpenAI** gateway returning schema-validated JSON with backoff retries on rate limits, and **BullMQ** workers moving TTS synthesis and S3 upload off the request path. **Tech:** NestJS, Prisma, Redis, BullMQ, AWS S3

🔗 **VidSense: Multimodal AI Video Intelligence Platform (FYP)** | *RAG, LLMs, Whisper, FastAPI*

- RAG pipeline over video transcripts using **sentence-transformers** embeddings with **FAISS/ChromaDB**, and cascading transcription (YouTube API to **Whisper Large-v3-Turbo**), achieving **92%** Q&A accuracy with timestamp-accurate clip retrieval.
- Hybrid algorithmic and LLM pipeline for highlight generation, AI podcasting, and automated dubbing/subtitling across **10+ languages**, cutting processing time **55%** via CUDA acceleration. **Tech:** Python, PyTorch, Ollama, FastAPI, FAISS, ChromaDB, FFmpeg

Education

Institute Of Business Administration (IBA)

June 2025

Bachelor of Science in Computer Science

Karachi, Pakistan

Technical Skills

AI/LLM: RAG, LangGraph, LangChain, Agentic AI, ReAct Agents, GPT-4o, Azure OpenAI, Prompt Engineering, OpenAI Embeddings, Whisper, Ollama

Vector & Retrieval: pgvector, FAISS, ChromaDB, Semantic Search, Chunking Strategies, Similarity Dedup

LLMOps & Observability: Langfuse, Arize Phoenix, LLM Tracing

Backend: Python, FastAPI, Celery, NestJS, Node.js, Spring Boot, REST APIs, SSE

Frontend: TypeScript, Next.js, React, Flutter, Tailwind CSS

Data & Infra: PostgreSQL, Redis, MongoDB, MySQL, Prisma, Docker, AWS S3, Git