

# MUZZAMIL KHALID

## AI ENGINEER — AGENTIC & GENERATIVE AI

Karachi, Pakistan | +92 370 308 8155 | [muzzamilkhalidai@gmail.com](mailto:muzzamilkhalidai@gmail.com)  
[muzzamilkhalid.com](https://muzzamilkhalid.com) | [github.com/Maimuzamilhu](https://github.com/Maimuzamilhu) | [linkedin.com/in/maimuzamil](https://linkedin.com/in/maimuzamil)

### SUMMARY

AI Engineer specializing in agentic and generative AI, building and shipping production LLM systems end to end — multi-agent orchestration (LangGraph, MCP), agentic RAG, real-time voice, and inference optimization, delivered as FastAPI services with Docker and CI/CD. Has shipped multiple live GenAI products, including a monetized SaaS serving 150+ daily active users and a distributed GPU pipeline that cut inference cost ~50%. Strong across LLM application development, RAG, agents and tool use, and LLMOps.

### EDUCATION

#### BS in Artificial Intelligence

Dawood University of Engineering & Technology (DUET), Karachi

Oct 2022 – Jun 2026

CGPA: 3.27 / 4.0

### TECHNICAL SKILLS

**Agentic AI & LLMs:** LangChain, OpenAI SDK, CrewAI, LangGraph, MCP, ReAct, tool/function calling, guardrails / prompt-injection defense

**LLM Foundations:** Transformer architecture, attention & self-attention, KV cache, PagedAttention, Grouped-Query Attention (GQA)

**RAG & Retrieval:** Agentic RAG, hybrid search (BM25 + vector), reranking, embeddings, Qdrant, Pinecone, RAGAS evals, Langfuse

**Inference & Optimization:** SLM fine-tuning & pre-training, tokenizer training, quantization, semantic caching, model routing, speculative decoding, Hugging Face

**Voice & Real-Time:** LiveKit, real-time STT/TTS, WebSockets, turn detection / endpointing (VAD), latency optimization

**Backend, Cloud & LLMops:** Python, FastAPI, SQL, Docker, Kubernetes, CI/CD (GitHub Actions), cloud (GCP · DigitalOcean · Cloudflare), observability (Prometheus/Grafana)

### SELECTED PROJECTS

#### CosMentor — Real-Time Voice AI Coding Tutor (Final-Year Project)

[cosmentor.app](https://cosmentor.app)

Multi-Agent Orchestration · LiveKit · Qdrant · React 19 · FastAPI · Neon PostgreSQL

- Built a real-time, two-way voice AI tutor that teaches live coding across Python & JavaScript through 40+ structured lessons, pairing spoken conversation with a shared in-browser IDE.
- Architected a multi-agent system (Orchestrator, Code Architect, Curriculum Specialist) with a LiveKit real-time voice layer and persistent long-term memory (Qdrant); cut conversational latency by optimizing turn-detection / endpointing (VAD) in the voice loop.

#### Realistic Handwriting — Text-to-Handwriting SaaS

[realistichandwriting.com](https://realistichandwriting.com)

Next.js · Firebase · Stripe/Polar · OCR · TTS

- Designed and shipped a live, monetized SaaS with 150+ daily active users and 80+ monthly subscribers, converting typed text, PDFs, and YouTube transcripts into realistic handwriting with a rich editor, OCR (Tesseract/scribe.js), and PDF export.

#### DreamTuber — AI Faceless-Video Autopilot

[dreamtuber.com](https://dreamtuber.com) · [Product Hunt](#)

Next.js · RunPod serverless GPU · LTX-Video · Chatterbox TTS · Cloudflare R2

- Built a multi-agent video "autopilot" that writes long-form scripts, clones narration voices (Chatterbox TTS), generates scene visuals and renders video (LTX-Video) on RunPod serverless GPU workers, then auto-publishes to YouTube; cut GPU inference cost ~50% via a Q8\_0 GGUF (int8) image model and a distilled LTX tier requiring fewer diffusion steps.

#### ClawStart — MCP Agent Runtime

[GitHub](#)

MCP · mem0 · ChromaDB · ReAct · TypeScript · Docker

- Built an agent runtime that connects external MCP servers and exposes 47 tools (GitHub 26 + Notion 21) into a ReAct loop (Think → Act → Observe), with a schema converter mapping tool definitions into validated arguments; split memory by lifetime — mem0 for durable facts across restarts, ChromaDB for vector retrieval over conversation history — delivered over Telegram, containerised.

### EXPERIENCE

#### AI Engineering Intern — Logitech Digital

Feb 2026 – Mar 2026

- Built and optimized production RAG pipelines (hybrid retrieval, chunking, embeddings) over enterprise knowledge bases; reduced query latency ~3s→1s via response caching, ANN index tuning, and top-k optimization.
- Collaborated in a cross-functional team and shared codebase to build LLM orchestration layers and agent-based systems, shipping features integrated via REST APIs.

#### EEG Data Analysis & Pattern Recognition Intern — Embedded Neuro Systems Lab, DUET

Jul 2025 – Aug 2025

- Engineered EEG signal-processing pipelines (ICA, FFT, digital filtering) for feature extraction and classification, improving neural-signal accuracy for the lab's ongoing research.

### CERTIFICATIONS

- MCP Protocols (Anthropic)
- LangGraph Foundation (LangChain Academy)
- Google Prompting Essentials
- Google Data Analytics Professional Certificate
- Reinforcement Learning from Scratch (Vizura)