

MINHAL ABDUL SAMI

Senior AI Engineer — Agentic Systems, Multimodal AI & Arabic-First LLM Infrastructure

Pakistan (Remote) • Secondary base: Jeddah, Saudi Arabia | romeomino415@gmail.com | +92 304 878 1849

linkedin.com/in/minhal-abdul-sami | github.com/Minhal-sami1 | github.com/Unisami

SUMMARY

AI engineer specializing in agentic LLM systems, multimodal (video + audio + text) pipelines, and Arabic-first bilingual AI for Gulf enterprise clients. Current production work: on-premise LLM infrastructure under Saudi/Qatar data-residency constraints, multi-agent architectures in AutoGen/LangGraph, LLM-as-Judge evaluation frameworks, and video content analysis pipelines. Comfortable owning a problem from modeling through containerized deployment and monitoring.

CORE STACK

Models/Frameworks: GPT-4/4o, Claude (Sonnet/Opus), Llama 3, Qwen, Whisper, open-source VLMs; LangChain, LangGraph, AutoGen, CrewAI, LlamaIndex, Haystack.

LLM Ops: Self-hosted inference (vLLM, Ollama), prompt caching, LLM-as-Judge and rubric-based evals, regression harnesses, structured outputs, RAG over Chroma/FAISS/pgvector, hybrid search (BM25 + dense).

Multimodal/Video: Self-hosted Whisper for ASR, MoviePy/ffmpeg for video pipelines, subtitle alignment across 14 languages, engagement-signal detection on long-form video.

Backend/Infra: Python (primary), Node/TypeScript, FastAPI, Flask, PostgreSQL, Redis, Docker, WebSockets, Azure (OpenAI, App Service, Functions), AWS basics.

Languages/Domains: Arabic-first bilingual UX (RTL), Gulf government & enterprise, legal-domain RAG, sales/CRM automation, content intelligence.

EXPERIENCE

Senior AI Engineer

Ghaia.ai — Remote (Doha, Qatar)

Oct 2025 – Present

AI consultancy shipping production LLM systems for Gulf enterprise clients: Qatar's National Planning Council (NPC), Qatar Airways, and Qatar's Ministry of Communications & IT (MCIT).

- **Multi-agent content platform (NPC):** architected the AI layer of a WebSocket-based AutoGen system generating Arabic-first government content across 10+ workflows (press releases, social captions, infographics, posters, video scripts). Own agent behavior, model routing, tool-call schemas, and session/state management.
- **PromptGraph engine:** designed a declarative, state-gated prompt composition engine replacing a 750-line imperative builder with ~24 independently testable sections. Cut in-session prompt size ~45% in review/approved states; added an explain(ctx) debugger so prompt behavior is inspectable.
- **Arabic-first bilingual pipelines:** enforced Qatari conventions (official nomenclature, honorifics, MSA register), Arabic-authoritative transcreation to English, and RTL-aware rendering across HTML/Playwright, PPTX, SVG, and PDF output paths.
- **Legacy orchestrator migration:** led a 7-phase cutover from a monolithic orchestrator to a workflow-module runtime. Deleted 13 tool-factory files (~220 LOC) with zero downstream contract breakage, gated by line-by-line split-accounting tests and two-stage shadow-execution equivalence runs.
- **E2E agentic test pipeline:** 41-test suite (~23 min on 4 workers) covering every workflow with tool-path / prompt-source / state-transition instrumentation, plus LLM-as-Judge scoring for Arabic outputs. Now the regression gate for all agent changes.
- **Evaluation infrastructure (Qatar Airways):** built LLM-as-Judge harnesses, golden datasets, and a date-parser evaluator for the travel-discovery pipeline. Authored a batch-API proposal collapsing 30 sequential fare-calendar calls into a single aggregated endpoint.
- **AI engineering methodology:** created and rolled out the L3 AI Workflow (PRD / ADR / CONVENTIONS / TASK_TEMPLATE / AGENTS.md) for structuring Claude Code and Copilot sessions across the team; submitted as a company "AI in Action" initiative.

AI Engineer

Unisami — Remote (Austin, TX, US)

Jun 2023 – Oct 2025

Early-stage US startup. I was the engineer on AI products; a co-founder owned go-to-market. Shipped five LLM/agentic products to production.

- **ACOS (sales automation):** multi-agent outreach system in AutoGen + LangGraph over Azure OpenAI. Lead enrichment, ICP scoring, email drafting, and reply classification wired into a React dashboard. Processed ~10K leads/month; internal reporting showed meaningful lift in qualified-lead rate, but I don't have an auditable before/after number.
- **Autonomous Social Media system:** CrewAI pipeline (strategist → writer → critic → scheduler) on PostgreSQL/SQLAlchemy. Generated and scheduled content across six platforms; operator time dropped from a full-time role to occasional review.
- **AI News Aggregator:** Python + Feedparser/BeautifulSoup ingest → GPT-based summarization and SEO rewrite → WordPress publishing. Grew from 0 to ~1.2K weekly organic readers in ~8 weeks with fully automated publishing.
- **AI Qanoon (Saudi legal RAG):** retrieval over ~1 GB of scraped Saudi legal documents (PDF + HTML). FAISS index, chunking tuned for clause boundaries, Arabic NLP preprocessing. First usable version shipped in ~1 week; spent several more weeks on answer grounding, citation rendering, and Arabic UI polish.
- **Viralizeit (video analysis):** ingest long-form video → self-hosted Whisper transcription across 14 languages → GPT-based engagement-segment detection → MoviePy clip generation. Pipeline produced ~2,800 short-form clips across client channels; aggregate reach in the multi-million range.

Senior Coding Instructor (part-time, concurrent)

Codingal — Remote

Jul 2022 – Feb 2025

- Taught Python, web, and intro-ML to 500+ students (K-12 and adult). Designed curriculum modules and graded programming competitions. Ran alongside full-time engineering roles.

Chief Technical Officer (short engagement)

Adrenod — Lahore, Pakistan

May 2023 – Aug 2023

- 3-month founder-stage engagement. Designed the initial backend and deployment architecture for their platform; left to focus on AI work at Unisami.

Head of IT (part-time, concurrent)

Ribat Store — Pakistan

Sep 2022 – Nov 2023

- Ran IT for a retail operation (~75 users): network, endpoint security, vendor negotiation, and internal tooling. Non-AI work; listed for completeness.

Software Engineer

Devaxl

Dec 2021 – May 2022

- Backend and integrations work. Restructured an expensive third-party API integration so a client could drop the paid tier entirely.

Co-Founder

Webspire — Remote

May 2021 – May 2022

- Small agency building websites and API backends for SMB clients. First real production work; shut it down to go deep on AI.

SELECTED PROJECTS

AI Qanoon — Saudi Legal AI Assistant

Python · LangChain · Azure OpenAI · FAISS · Arabic NLP · Flask

- RAG system over ~1 GB of Saudi legal corpus with clause-aware chunking, hybrid retrieval, and citation-first answers. Public, no-login UX.
- Handled end-to-end: scraping → preprocessing → embeddings → retrieval → UI. First working version in one week; ongoing tuning afterward on grounding and Arabic display.

ACOS — Automated Customer Outreach System

Python · React · Azure OpenAI · LangChain · LangGraph · AutoGen

- Multi-agent sales pipeline: lead enrichment, qualification, personalized outreach, and reply classification, with role-based multi-user access.
- Processed ~10K leads/month; clear qualitative improvement in rep efficiency, though I do not have a clean auditable lift metric I'd defend in an interview.

Viralizeit — Video Content Intelligence

Python · self-hosted Whisper · OpenAI · MoviePy · Flask

- Long-form video → self-hosted Whisper transcription (14 langs) → engagement-segment scoring → automatic short-form clip generation.
- Closest project to a computer-vision/video-analysis role: heavy work on audio transcription, segment ranking, and video-processing pipelines; light CV beyond scene cuts.

Autonomous Social Media System

Python · CrewAI · Azure OpenAI · PostgreSQL · SQLAlchemy

- Multi-agent content pipeline (strategist/writer/critic/scheduler) publishing to six platforms with feedback loop from engagement data.

AI News Aggregator

Python · OpenAI · Flask · BeautifulSoup · Feedparser · WordPress

- Fully autonomous ingest → summarize → SEO-rewrite → publish loop. Grew to ~1.2K weekly organic readers with no human in the publishing path.

CERTIFICATIONS

- **Neo4j Certified Professional** — Neo4j (Oct 2025)
- **CS50: Introduction to Computer Science** — HarvardX (Dec 2020)
- **CS50 Web Programming with Python and JavaScript** — Harvard (Aug 2022)

EDUCATION & BACKGROUND

BSc Computer Science (in progress) — University of the People

Admitted Jul 2025, expected Jun 2029 (part-time alongside full-time engineering).

Primarily self-taught. Started building production software in 2021 (Webspire), moved into AI engineering at Unisami in 2023, and specialized into agentic LLM systems at Ghaia.ai in 2025. Happy to discuss trajectory directly in an interview.