

MUHAMMAD HAARIS

AI Engineer · Production LLM Systems · Healthcare AI

muhammad.haaris2003@gmail.com | +92-304-944-4702 | linkedin.com/in/haaris278 | github.com/MuhammadHaaris278 | Lahore, Pakistan

PROFESSIONAL SUMMARY

AI Engineer specializing in production LLM systems, agentic workflows, and healthcare AI infrastructure. Experience building and deploying clinical RAG and NLP systems for healthcare providers with focus on reliability, low hallucination rates, scalable backend architecture, and observability. Skilled in LangGraph orchestration, hybrid retrieval systems, FastAPI services, and HIPAA-compliant AI pipelines. Comfortable working in async remote environments with end-to-end ownership.

TECHNICAL SKILLS

AI / ML	LangChain, LangGraph, HuggingFace Transformers, PyTorch, OpenAI APIs, Gemini APIs, Sentence-Transformers, Cross-Encoders, PubMedBERT, SciSpaCy, MedSpaCy
RAG Systems	Qdrant, FAISS, Hybrid Search, BM25, Reciprocal Rank Fusion (RRF), MMR Reranking
Backend	Python, FastAPI, PostgreSQL, Redis, REST APIs, JWT Authentication, Async Programming
DevOps	Docker, Docker Compose, GitHub Actions (CI/CD), Prometheus, Grafana
Healthcare AI	FHIR R4, HL7, ICD-10, CPT, RxNorm, LOINC, Clinical NER, HIPAA Compliance
Languages	Python, SQL, JavaScript / TypeScript, C++

PROFESSIONAL EXPERIENCE

AI Engineer

Optikode

Aug 2025 – Present · Lahore, Pakistan

- Architected and deployed MedGraph AI, a production clinical RAG assistant used in healthcare provider workflows.
- Built AI Clinical Scribe for extracting ICD-10, CPT, RxNorm, and LOINC entities from clinical documentation.
- Designed fault-tolerant multi-LLM orchestration pipelines with automated retry and fallback handling, maintaining 99%+ uptime.
- Developed containerized FastAPI services with GitHub Actions CI/CD and automated testing pipelines.
- Implemented HIPAA-oriented audit logging, patient data isolation, and monitoring with Prometheus and Grafana.

AI Engineer Intern

Kryptomind

Jun 2025 – Aug 2025 · Lahore, Pakistan

- Built RAG pipelines using OpenAI, FAISS, and Qdrant with sub-second retrieval latency on 50K+ documents.
- Improved retrieval precision by 22% through embedding evaluation and retrieval optimization.
- Developed LangGraph agents with session memory and streaming responses using FastAPI.

PRODUCTION PROJECTS

MedGraph AI — Clinical Intelligence Platform

LangGraph, FastAPI, Qdrant, Redis, PostgreSQL, OpenAI, Gemini, Docker

- Built a hybrid retrieval and reranking pipeline that reduced response latency by approximately 40%.
- Added grounding and source verification checks to reduce hallucinations before clinician delivery.
- Identified and fixed a critical cross-patient data leakage issue prior to production deployment.
- Improved benchmark accuracy from 46% to production-ready performance through retrieval and prompt optimization.

AI Clinical Scribe — Medical Coding Engine

FastAPI, SciSpaCy, MedSpaCy, FHIR R4, Docker

- Developed NLP pipelines for extracting structured medical codes from clinical notes.
- Designed validation workflows to reduce false-positive CPT predictions.
- Integrated healthcare ontologies and FHIR-compatible adapters for clinical interoperability.

AgentGraph SQL Agent — Multi-Agent AI System

LangGraph, FastAPI, SQL, Tool Calling, Observability

- Built an agentic workflow for multi-database querying, reasoning, and execution.
- Added token-cost tracking and observability tooling for debugging and optimization.
- Designed reusable infrastructure adaptable beyond healthcare use cases.

ADDITIONAL PROJECTS

- **CGM Interpreter** — Time-series ML platform for continuous glucose monitor analysis and visualization.
- **VLM Fine-Tuning** — Fine-tuned BLIP-2 and CLIP models for medical image captioning and multimodal understanding.

EDUCATION

Bachelor of Computer Science <i>FAST – National University of Computer and Emerging Sciences</i>	2021 – 2025 · Lahore, Pakistan
Intermediate (Pre-Engineering) <i>Punjab College</i>	2019 – 2021 · Lahore, Pakistan

Project repositories and references available upon request