

MUHAMMAD TALHA KHAN

+92 332-4755831 ✦ Lahore, Punjab

muhammadtalhakh006@gmail.com ✦ [GitHub](#) ✦ [Portfolio](#) ✦ [LinkedIn](#)

SUMMARY

ML & AI Engineer with 3+ years of experience designing intelligent systems and shipping production AI products in healthcare and SaaS. Combines hands-on expertise in LLMs, RAG, and MLOps with team leadership and a track record of turning AI prototypes into scalable, real-world solutions.

EDUCATION

Bachelor of Software Engineering, COMSATS University Islamabad, Lahore Campus 2021 – 2025
CGPA: 3.89/4.00 — *Gold Medalist*

SKILLS

Core ML/AI	Python, Traditional ML, Deep Learning, Fine-Tuning (LoRA, QLoRA), RAG, PageIndex (Vectorless RAG), Vector Databases
AI Frameworks	LangGraph, CrewAI, OpenAI Agents SDK, AutoGen, Parlant, FastAPI
Platforms	Hugging Face, Ollama, Retell AI, Kaggle, Xano, LiveKit, Opik
Full-Stack	MERN, MEAN, Next.js, Nest.js, React Native, Swift, Java, SQL, NoSQL
Cloud & MLOps	AWS (SageMaker, Bedrock), Docker, Kubernetes, NGINX, MLflow, Git
Automation	Playwright, Selenium, Scraping, Zapier, MAKE

EXPERIENCE

ML Engineer Sept 2025 - Current
Persivia *Lahore, Punjab*

- Deployed scalable infrastructure with NGINX load balancing, Docker Swarm orchestration, and Jenkins CI/CD pipelines for high availability.
- Implemented a RAG pipeline with Milvus vector DB, improving AI accuracy and semantic search over large-scale medical knowledge bases.
- Shipped LLM-based tools (Ambient Scribe, CareTalk, QueryBuilder, FHIRScribe) for clinical documentation, natural language data queries, and FHIR standardization.

Lead Software Engineer Sept 2023 - Sept 2025
KodeXLabs *Lahore, Punjab*

- Led 4-6 engineers across SaaS products, owning architecture and sprint planning.
- Redesigned core workflows and refactored legacy code, reducing response times by nearly 40%.
- Set up CI/CD pipelines and code review standards, speeding up deployments and cutting manual work.

Software Engineering Intern June 2023 - Sept 2023
TechQode *Lahore, Punjab*

- Improved frontend performance by optimizing components and reducing bundle size.
- Integrated REST APIs to streamline data flow and enhance responsiveness.
- Built and deployed a full-stack Next.js app using Docker on DigitalOcean.

PROJECTS

Ambient Scribe. AI-powered clinical documentation system using NLP and Speech AI to transcribe doctor-patient conversations into structured medical notes, reducing manual data entry for healthcare providers.

CareTalk. Natural language interface for patient-level data retrieval, converting clinician queries into structured database calls to surface individual patient records and history in real time.

QueryBuilder. Population-level analytics tool that translates natural language prompts into complex database queries, enabling clinicians to explore cohort trends and aggregate insights without writing SQL.

FHIRScribe. Pipeline that converts raw clinical documents into FHIR-compliant JSON at scale (**20,000+** documents), enabling seamless ingestion into EHR systems.

Genaiva. Led development of a real-time AI voice agent system for automated restaurant order-taking, built with LLMs, voice AI, and real-time streaming. ([Live](#))

FundedFlow. Marketing and trading platform built with Next.js, Supabase, Brevo, and Sanity, enabling user acquisition, campaign automation, and seamless trade management. ([Live](#))

CERTIFICATIONS

- **AWS Certified AI Practitioner** — Amazon Web Services
- **AWS Certified GenAI Architect** — Amazon Web Services
- **AWS Certified Solutions Architect** — Amazon Web Services
- **HIPAA Compliance Certification** — North American Institute
- **Model Context Protocol (MCP)** — Anthropic
- **Machine Learning** — SAMZ Technologies

LEADERSHIP & ACTIVITIES

- Led small engineering teams through full product cycles, from planning and architecture to shipping and iteration.
- Mentor junior engineers on ML, RAG, and building real-world AI systems beyond tutorials.
- Into traveling, cricket, and strategy games outside of work.