

Syed Muhammad Mujtaba

AI Engineer — LLM Applications, Agentic Systems & RAG | Full-Stack

Lahore, Pakistan | +92 303 4053767 | mujtabashah1204@gmail.com

linkedin.com/in/mujtaba-shah | github.com/Mujtabashah4

PROFESSIONAL SUMMARY

AI Engineer with 3+ years of experience building and shipping production LLM applications for enterprise clients. Experienced across the full GenAI stack — agentic systems, RAG pipelines, multi-agent workflows, prompt engineering, MCP integrations, and real-time voice AI. Works AI-native, building daily with Claude, GPT, and agentic coding tools such as Claude Code and Cursor. Full-stack engineer across TypeScript/Node.js, React, and Python, with an MS in Artificial Intelligence from LUMS.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, Java, C/C++, SQL

Generative AI & LLMs: Claude (Anthropic API), GPT-4o / Realtime API, Mistral, LangGraph, LangChain, Agentic AI & Multi-Agent Workflows, Tool / Function Calling, Retrieval-Augmented Generation (RAG), Prompt Engineering, LLM Evals & Guardrails, Fine-Tuning & Quantization, Model Context Protocol (MCP), Hugging Face Transformers, PyTorch, TensorFlow

AI-Assisted Engineering: Claude Code, Cursor, GitHub Copilot, v0.dev, AI Code Review

Retrieval & Vector Search: FAISS, Pinecone, Qdrant, Hybrid Retrieval (BM25 + Dense), Cross-Encoder Reranking, Query Rewriting, Semantic Search

Backend & APIs: Node.js, Express.js, FastAPI, Flask, REST APIs, WebSockets, BullMQ, Mongoose, GitHub API (Octokit), Microservices, Event-Driven Architecture

Frontend: React.js, Next.js, Vite, Tailwind CSS

Cloud & DevOps: AWS, Docker, Kubernetes, CI/CD, GitOps (ArgoCD)

Databases & Messaging: MongoDB, PostgreSQL, Redis, SQLite, Apache Kafka

Voice AI: Telnyx Voice API, Genesys, Twilio, Retell, Vapi

PROFESSIONAL EXPERIENCE

AI Innovation Specialist – Software Engineering

Mar 2026 – Jul 2026

IgniteTech (Crossover) — Remote

- Built and shipped a customer-onboarding automation platform for an enterprise GenAI product — owned end-to-end from concept to live enterprise customer rollouts (Node.js/TypeScript, React, MongoDB, AWS, Kubernetes, GitOps/ArgoCD).
- Worked on IgniteTech's AI innovation team, embedding Generative AI across enterprise software products and modernizing how software is designed, built, and delivered.
- Drove AI-native engineering across the team — daily agentic development with Claude Code, Model Context Protocol (MCP) integrations with enterprise platforms, and hands-on evaluation of LLMs (Claude, GPT, Mistral) to accelerate delivery and raise code quality.
- Partnered directly with enterprise architects, SRE, and design to translate customer requirements into shipped features, iterating from working prototype to production.

AI Engineer

Jul 2025 – Nov 2025

HytGenx

- Built retrieval-augmented generation (RAG) pipelines grounding LLM outputs with domain-specific knowledge bases, achieving 85%+ answer relevance on internal evaluation benchmarks.
- Fine-tuned transformer-based language models on task-specific datasets, reducing inference latency by 35% through model quantization and prompt optimization while maintaining output quality.
- Automated knowledge-intensive internal workflows using LLM-based agents, cutting manual processing time by approximately 50% across repetitive document analysis tasks.

AI/ML & Full-Stack Developer

Oct 2024 – May 2025

Splendid Mark

- Developed intelligent dashboards integrating LangChain agents and LLMs to deliver NLP-powered analytics for market performance analysis, improving reporting turnaround by 40%.
- Built end-to-end data pipelines collecting and processing data from Instagram Graph API, Facebook Marketing API, Reddit, and web sources via BeautifulSoup for real-time trend analysis and competitor monitoring.
- Implemented vector search and RAG workflows using Pinecone, cutting retrieval latency from 320 ms to 95 ms and improving relevance scoring in AI-driven analytics systems.

Software Engineer / Technical Instructor

Jul 2023 – Mar 2025

Knowledge Streams

- Led MERN stack and software fundamentals training (MongoDB, Express, React, Node) and taught OOP, DSA, and database design, mentoring cohorts to deliver production-ready projects.
- Guided students through API design, deployment workflows, and best practices for version control and CI/CD.

Web Developer

Dec 2022 – Apr 2023

Hazel Mobile

- Developed interactive web applications using JavaScript and React.js, contributing to backend API integration connecting frontend components with databases and server-side logic.

PROJECTS

Agentic RAG System with Self-Correcting Retrieval — *LangGraph, GPT-4o, Qdrant, Python*

- Designed an autonomous retrieval agent that grades its own search results, rewrites failing queries, and falls back to live web search — built as a cyclic LangGraph state machine with hallucination and grounding checks before every response.
- Lifted top-5 retrieval precision from 72% to 89% through hybrid retrieval and cross-encoder reranking.

Enterprise AI Content Generation Platform — *LangGraph, GPT-4o, Claude, FastAPI, Next.js, Google Cloud*

- Built a multi-agent platform that autonomously produces personalized course materials — slides, documents, curricula, and assessments — adapted to each instructor's teaching style; deployed end-to-end on Google Cloud with a Next.js frontend and FastAPI backend.

Real-Time Conversational Voice AI — *Python, Telnyx, Genesys, Retell, Vapi*

- Engineered a live voice-agent platform handling inbound phone calls with LLM-driven dialogue at under 500 ms end-to-end latency, including speech synthesis, transcription, and dynamic agent provisioning.

EDUCATION & CERTIFICATIONS

Lahore University of Management Sciences (LUMS)

Sep 2024 – Jun 2026

MS in Artificial Intelligence

- Coursework: Large Language Models, Data Science & Visualization, Machine Learning. Research focus: advanced retrieval strategies for RAG systems, LLM evaluation and grounding methods.

Government College University (GCU), Lahore

Sep 2019 – Jun 2023

BS in Computer Science — GPA 3.45/4.0

- Certification: MERN Stack Development — Knowledge Streams (Jul 2023 – Oct 2023).