Muhammad Muaz Ashraf

AI Engineer | Generative AI & Agentic Systems

Lahore, Pakistan | +92 323-8479018 | **Email | LinkedIn | GitHub | Portfolio**

Summary

AI Engineer with 3+ years of hands-on experience building production-ready Generative AI systems, including RAG pipelines, agentic architectures, MCP servers, and multimodal AI workflows. Strong expertise in Python-based backend development, LLM integration, retrieval optimization, and AI automation, with a proven track record of delivering measurable performance and productivity gains.

Experience

AI Engineer (Generative AI & NLP) — Mojo Solo, Remote

Mar 2023 – Present

- Developed an automated Slack bot that instantly deploys GitHub repositories as web services, CLI, or TUI tools.
- Built an automated CI pipeline that receives bug/feature reports via webhook, routes them to Claude Code on a local device, and auto-generates PRs or direct commits to production repos.
- Built and deployed advanced RAG pipelines using LangChain, LlamaIndex, and vector databases, improving answer relevance by ~80% through hybrid retrieval, hashing, chunking and cross-encoder reranking.
- Designed agentic AI systems with LangGraph, CrewAI, and MCP-based architectures, enabling multi-step reasoning, tool execution, and long-term memory handling.
- Developed custom MCP servers to support scalable agent workflows, memory orchestration, and external tool integration.
- Engineered voice-based AI agents for automated appointment booking using Retell AI, Twilio, LiveKit, Deepgram, and ElevenLabs.
- Built multimodal AI pipelines combining text, image, audio, and video generation and processing.
- Created AI video generation workflows integrating LLMs, image models, TTS, and video composition tools, reducing manual editing time by ~70%.
- Optimized LLM latency, cost, and prompt reliability for production deployment.
- Implemented activity detection systems for fitness applications (pushups/squats counting) and face authorization using object detection and OpenCV computer vision and Yolo.

Tech: Python(Fast API, Flask), Generative AI, LLMs, RAG, LangChain, LangGraph, LlamaIndex, MCP, Flow, Pinecone, Qdrant, Chromadb, Twilio, LiveKit, Deepgram, ElevenLabs, Hugging Face, COLPALI, Claude-code, Multimodal AI, Docker

Education

BSc in Software Engineering | University of Lahore | 2018 - 2023

Certification

- Event-Driven Agentic Document Workflows, 2025
- Building AI Voice Agents for Production, 2025
- Multimodal RAG: Chat with Videos!, 2024

Languages

English (Professional) · Urdu (Native)