

Muzammil Nawaz Khan

[+92 304 6395539](tel:+923046395539) | [Gmail](#) | [Linkedin](#) | [Github](#) | [Portfolio](#)

Professional Summary

Results-driven AI Engineer with 1+ years of experience in Generative AI, Voice Agents, and Computer Vision. Proven track record of delivering production-grade AI solutions, including optimizing voice agent workflows that reduced front-desk handling time by 40% and lowered inference costs by 35%. Expert in deploying autonomous agentic frameworks and high-accuracy deep learning models, with an award-winning research background in biomedical segmentation. Passionate about architecting scalable, cost-efficient automation to solve complex enterprise challenges.

Technical Skills

- **AI & Machine Learning:** PyTorch, TensorFlow, Keras, Scikit-learn, OpenAI API, LLMs, LangChain, LangGraph.
- **Voice & Agents:** LiveKit, Silero VAD, ElevenLabs, n8n (MCP Server), STT/TTS.
- **Backend & Web:** FastAPI, Next.js, WebSockets, SSE.
- **Computer Vision:** OpenCV, U-Net, SegFormer, YOLO, Object Detection.
- **Languages & Tools:** Python, C++, SQL, Git, Docker, NumPy, Pandas.

Professional Experience

CareCloud | AI Engineer

July 2025 – Present

- Architected a real-time Voice AI agent using LiveKit and ElevenLabs, integrating Silero VAD for low-latency interruptions, reducing front-desk handling time by 40%.
- Designed an n8n-based orchestration layer acting as an MCP server to manage state and logic-based routing, decreasing inference costs by 35% through efficient prompt caching.
- **Agent Orchestration:** Implemented an n8n workflow acting as a Model Context Protocol (MCP) server to manage conversation state, backend prompts, and call routing logic.
- **Data Pipeline Automation:** Engineered Python pipelines to interface with Hikvision NVR systems to fetch raw data, clean it, and dump it into a database.

Risotech | AI Research Intern

July 2024 – Aug 2024 | Islamabad, Pakistan

- **Neurofusion Project (FYP):** Spearheaded the development of a 3D brain tumor segmentation system using the BraTS20 dataset.
- **Model Optimization:** Benchmarked 3D U-Net against SegFormer3D, selecting the latter to achieve 85%+ accuracy.
- **Awards:** Won 1st Place at COMPPEC 2025 and received the Departmental Silver Medal

for technical complexity.

Projects

AI Agents & Generative AI

Conversational AI Agent (ReAct)

LangGraph, FastAPI, Next.js, OpenAI

- Developed a ReAct-style AI agent featuring tool calling, visible thought processes, and real-time WebSocket streaming.
- Integrated six custom tools, including Calculator, Weather, DateTime, Web Search, Code Interpreter, and RAG Search.
- Engineered a persistent memory system using a SQLite checkpointer to maintain session state across multi-turn conversations.

Document Summarizer

FastAPI, Next.js 14, LangChain, ChromaDB

- Engineered a production-grade multi-strategy document summarization platform with an automatic strategy router (Stuff, Map-Reduce, Refine, Recursive).
- Implemented live SSE streaming for real-time token delivery and structured output generation.
- Engineered a production-grade multi-strategy summarization platform; implemented asynchronous processing pipelines that reduced average document analysis latency by ~50% while maintaining 99% factual consistency in RAG-based extraction.

RAG Custom Engine

Pure Python, GPT-4o-mini

- Built a Retrieval-Augmented Generation pipeline entirely from scratch without external frameworks, implementing a custom HNSW vector store and Okapi BM25 keyword index.
- Designed a hybrid retrieval system using Reciprocal Rank Fusion (RRF) and weighted score ensembles.
- Integrated a 3-stage Self-RAG evaluation gate (retrieval decision, relevance grading, hallucination check) and contextual compression.
- Built a hybrid RAG pipeline from scratch, integrating HNSW vector storage and RRF; improved retrieval precision by 20% compared to standard dense-vector baselines through the implementation of a 3-stage self-evaluation gate.

RAG LangChain Chroma

LangChain, ChromaDB, FastAPI

- Developed a hybrid RAG application featuring dual-layer memory (in-chat session memory and cross-chat LLM-summarized memory).

Computer Vision

Retinal Image Segmentation

Python, OpenCV

- Developed multi-level thresholding and morphological segmentation methods to automate vessel visibility enhancement for medical analysis.

Optical Braille Recognition

Python, OpenCV

- Built a dot-pattern recognition system using noise filtering and contour analysis to translate Braille characters with high precision.

Skin Image Segmentation

Python, OpenCV

- Designed a skin lesion segmentation pipeline using Connected Component Labeling (CCL) and post-processing for clinical diagnostic support.

Education

National University of Sciences and Technology (NUST)

Bachelor of Computer Engineering | Nov. 2021 – July 2025

- Islamabad, Pakistan | CGPA: 3.25
- Relevant Coursework: Digital Image Processing, AI and Decision Support System, Computer Vision

Ghazali Premier College

F.Sc. Pre-Engineering | June 2018 – Aug 2020

- Lahore, Pakistan

Professional Certifications

- IBM AI Engineering Professional Certificate (Deep Learning, PyTorch, Keras)
- IBM Generative AI Engineering Professional Certificate (LLMs, Transformers)
- IBM RAG & Agentic AI Professional Certificate (LangChain, LangGraph)
- LLMOps (Large Language Model Operations) Specialization – Duke University
- Informed Clinical Decision Making using Deep Learning – University of Glasgow
- Modern Data Strategy for Enterprise GenAI – Fractal
- Complete Modern C++ (C++11/14/17) – Packt