Altaf Rehman

Islamabad, Pakistan — altafrehman.dev@gmail.com — LinkedIn

Professional Summary

AI Engineer with 3+ years of software engineering experience and over a year building production LLM applications, agent workflows and tool-integrated systems. Designed and deployed an end-to-end, in-house voice AI platform (LiveKit/WebRTC) supporting huge number of concurrent sessions at 300–500 ms latency, replacing commercial third-party solutions and reducing costs by 60%. Built BrainCX's Model Context Protocol (MCP) server and custom connectors for AI marketplaces including Claude and OpenAI, plus an in-house RAG knowledge base with hybrid search. Experienced with LangGraph, tool calling, structured outputs and iterative model evaluation. Author of two technical publications on low-latency voice agent architecture and multi-session reliability. Seeking to build dependable, permission-aware AI agents for public-sector use cases.

Education

B.S. Software Engineering — FAST National University of Computer and Emerging Sciences (NUCES), Islamabad

Relevant coursework: Machine Learning, Natural Language Processing, Distributed Systems, Operating Systems, Database Systems

Publications

- **"Architecture Patterns for Low-Latency Real-Time Voice Agents"**
  Zenodo | DOI: 10.5281/zenodo.23216507
- **"Concurrency and Reliability Challenges in Multi-Session Conversational AI Systems"**
  Zenodo | DOI: 10.5281/zenodo.23222718

Final Year Project

Real-Time Multi-Agent Conversational Voice Platform

Designed and implemented a complete real-time conversational voice system for multi-turn, task-oriented dialogues with low latency. The system integrated speech-to-text, LLM-based dialogue management and text-to-speech in a streaming pipeline. Key contributions included custom turn-taking logic, concurrent session management, and a modular agent architecture supporting both open-domain and goal-oriented conversations. The project emphasized production-level concerns such as latency optimization, reliability under concurrent load, and integration of modern LLM APIs with real-time audio transport.

Professional Experience

AI Engineer August 2025 – Present

BrainCX (Florida, USA) — Remote

- Designed, implemented and deployed a full production AI agent platform using LiveKit (WebRTC) for real-time audio, with LLM-driven dialogue management and custom prompt engineering. The system supports 60 concurrent sessions at 300–500 ms latency and replaced a commercial third-party vendor, delivering a 60% cost reduction while improving reliability.
- Developed the company's MCP (Model Context Protocol) server and custom connectors for AI marketplaces, including Claude and OpenAI, exposing platform capabilities as tools that external AI assistants can call.
- Built agent workflows with LangGraph and an in-house RAG knowledge base with hybrid search, enabling agents to ground responses in company data.
- Built end-to-end conversational agent infrastructure covering low-latency turn-taking, STT/TTS integration, tool and API integrations, and scalable backend services. Focused on production constraints including concurrency, latency and response quality under real user load.
- Owned platform backend, API integrations and iterative model evaluation. Migrated all company clients from the external vendor to the in-house system and continued shipping scalability and capability improvements.
- Developed additional conversational agents (e.g., hands-free web-form completion via natural language intent mapping), demonstrating applied dialogue systems beyond pure telephony.

MERN Full Stack Developer October 2024 – August 2025

Devoura — Remote (Islamabad)

- Built and maintained scalable full-stack applications (React, Next.js, Node.js, Express) with REST/GraphQL APIs that later supported AI feature extensions.
- Managed data stores (MongoDB, PostgreSQL, Redis) and established AWS, Docker, and GitHub Actions CI/CD pipelines with automated testing.

Back End Developer August 2023 – October 2024
Outsourcing Company — Remote (Part-time)

- Designed and maintained backend systems (Node.js, Express, MongoDB/PostgreSQL) exposing secure REST APIs, including early support for AI-powered product features.

Selected Projects

- LLM-based agent workflows (n8n) for multi-platform content generation and automated posting.
- End-to-end lead generation and personalized outreach pipeline combining structured data enrichment with LLM-generated messaging.
- Google Meet bot that joins meetings, records audio, and produces LLM-generated transcripts and summaries.

Technical Skills

Agents & Tool Integration: LangGraph, MCP Servers & Custom Connectors, Tool Calling, Structured Outputs, Agent Workflows, LLM Dialogue Management, Prompt Engineering
Voice & Conversational AI: LiveKit, Real-time Voice Agents, STT/TTS Integration, Low-latency Systems
AI / LLM & Retrieval: OpenAI-class APIs, RAG Pipelines, Hybrid Search, Vector Databases, Embeddings, Model Evaluation & Iteration
Languages & Backend: Python, FastAPI, TypeScript, Node.js, Express, REST, GraphQL, Socket.io, API Design
Frontend: React, Next.js, Tailwind CSS
Data & Infra: MongoDB, PostgreSQL, Redis, Vector DBs, AWS, Docker, GitHub Actions, CI/CD, Git
Other: Performance Optimization, Scalability & Concurrency