

Nouman Zahid

Islamabad, Pakistan | +92 345 3391928 | noumanzahi7676@gmail.com
https://linkedin.com/in/nouman-zahid | https://github.com/NoumanZahid-85

Summary

Full Stack AI Engineer with AI/ML and AWS cloud internship experience, building real-time voice AI, LLM agents and RAG systems with FastAPI, Next.js, LangGraph and WebSockets. Cut production voice latency from 3.0–3.7 s to 1.6–2.1 s and shipped 3 live AI products, including a RAG hallucination detector with 92.8% accuracy.

Technical Skills

Languages: Python, Java, JavaScript, TypeScript, SQL

AI and LLMs: LangChain, LangGraph, CrewAI, RAG, AI Agents, Prompt Engineering, MCP, PyTorch

Voice AI: Gemini Live (Speech-to-Speech), GPT-Live-1, Vapi, Whisper, ElevenLabs, LiveKit, Pipecat

Full Stack: FastAPI, Next.js, React, Node.js, Tailwind CSS, WebSockets, REST APIs

Data, Cloud and DevOps: PostgreSQL, Redis, SQLite, Qdrant, AWS, Docker, GitHub Actions, pytest, Git

Experience

AI/ML Engineer Intern

Jul 2026 – Aug 2026

Synctom

Pakistan

- Architected Haven: Real-time voice AI streaming 16 kHz audio over duplex FastAPI WebSockets to Next.js
- Engineered 2 voice engines: Gemini Live speech-to-speech and a Deepgram Flux, Groq and Aura-2 pipeline
- Reduced first-audio latency from 3.0–3.7 s to 1.6–2.1 s in production; shipped on Docker, PostgreSQL and Redis

AWS Cloud Engineering Intern

Jun 2025 – Aug 2025

CloudSec Network

Canada (Remote)

- Built and secured highly available AWS infrastructure (EC2, S3, Lambda, VPC) for production workloads
- Automated multi-tier deployments via Terraform (IaC), reducing provisioning time by 80% and config drift
- Hardened security with least-privilege IAM roles, restricted S3 policies, and CloudWatch monitoring

Projects

Haven | *FastAPI, WebSockets, Deepgram Flux, Groq, Gemini Live, PostgreSQL, Redis, Docker*

Live | GitHub

- Launched LLM/TTS on Flux's eager end-of-turn signal (50 ms audio latency) and ran RAG over partial transcripts, cutting planning time from 616 ms to 1 ms
- Designed an interruption judge that ignores echo/backchannels but stops on real interruptions in 0.84 s
- Implemented an LLM-free crisis circuit (0.85 s cached response), validated by 78 pytest tests and 17 red-team cases

TradeCraft HR | *FastAPI, LangGraph, Groq, PyGithub, SQLite, Next.js 15, FastMCP, ReportLab*

Live | GitHub

- Analyzed candidate repos via PyGithub to generate 12 tailored questions (3+ citing commits), gated by recruiter approval
- Managed interview state in SQLite and routed bluffing answers to follow-ups (capped at 2)
- Validated evidence-based scoring (rejecting scores without transcript quotes); exported ReportLab PDFs and FastMCP tools

Hallucinot | *Next.js 15, TypeScript, Jev/OpenJev, Tailwind CSS, Vercel*

Live | GitHub

- Transformed RAG answers into claims and scored sentence-passage pairs (72–81 checks) in 2.5–3.3 s using Jev probabilities
- Achieved 92.8% accuracy (26/28 sentences) at \$0.0006 per run, routing misses to human review
- Created a draggable review threshold to re-sort claims client-side with zero extra API calls

Education

COMSATS University Islamabad

Islamabad, Pakistan

Bachelor of Science in Computer Science

Sep 2023 – Jun 2027

- Completed coursework in Artificial Intelligence, Machine Learning, Database Systems, and Web Technologies